Analyse de contenus visuels en 2D et en 3D : évaluation de la pertinence d'un point de vue d'un objet 3D

Marie Pelissier-Combescure

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à Toulouse INP

Analyse de contenus visuels en 2D et en 3D : évaluation de la pertinence d'un point de vue d'un objet 3D

Thèse présentée et soutenue, le 24 juin 2024 par

Marie PELISSIER-COMBESCURE

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Informatique et Télécommunications

Unité de recherche
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Géraldine MORIN et Sylvie CHAMBON

Composition du jury
M. Jean-Luc MARI, Président, Aix-Marseille Université
M. Romain RAFFIN, Rapporteur, Université de Bourgogne
Mme Michèle GOUIFFES, Rapporteuse, Université Paris-Saclay
Mme Géraldine MORIN, Directrice de thèse, Toulouse INP
Mme Sylvie CHAMBON, Co-directrice de thèse, Toulouse INP

Titre : Analyse de contenus visuels en 2D et en 3D : évaluation de la pertinence d'un point de vue d'un objet 3D

Mots clés : Modèles 3D, cartes de profondeur, images 2D, qualité d'une représentation, saillance 3D, saillance curviligne, score de pertinence, réseaux de neurones, score de confiance, sélection de la meilleure vue, étude utilisateur et utilisatrice.

Résumé :
Dans notre monde tridimensionnel, les objets sont souvent représentés en deux dimensions, principalement visualisés à travers des supports 2D tels que des catalogues en papier ou des écrans d'ordinateur. La sélection du point de vue pour observer un objet en 3D sur un support 2D a un impact significatif sur son identification, la visualisation de ses caractéristiques et la compréhension de son utilité. Cette tâche est primordiale dans différents domaines applicatifs tels que les jeux vidéos, l'imagerie médicale, l'architecture ou la conception industrielle.

Cette thèse se penche sur la problématique du choix du point de vue 2D le plus adapté pour un objet 3D donné, avec pour objectif de mesurer la pertinence de ce point de vue au regard des « informations essentielles » de l'objet, c'est-à-dire les informations qui permettent de le reconnaître et d'en extraire les attributs caractéristiques. Nous proposons une méthode pour déterminer automatiquement le point de vue 2D le plus pertinent en extrayant et en quantifiant les « informations essentielles » de chaque point de vue disponible.

Deux axes de recherche ont été explorés dans cette étude. Le premier axe concerne l'évaluation de la pertinence des points de vue fixes offerts par des images texturées par rapport à un objet 3D. La géométrie de cet objet est définie par un modèle 3D. Le deuxième axe se concentre sur l'analyse d'un modèle 3D géométrique défini par un maillage 3D non texturé de l'objet. Nous proposons de déterminer automatiquement le point de vue le plus représentatif de l'objet, celui qui permet une identification et une compréhension sans ambiguïté.

Pour notre premier axe de recherche, nous développons une approche s'appuyant sur la détection des caractéristiques saillantes de l'objet tout en filtrant les informations non pertinentes comme celles liées à l'apparence, comme la texture, en tirant parti à la fois des images 2D et des modèles 3D. Nous introduisons un score de pertinence, dérivé des attributs géométriques et des recommandations liées à la photographie, qui nous permet de classer tous les points de vue disponibles. Pour valider l'approche, nous comparons les scores de pertinence avec les scores de confiance obtenus à partir de méthodes d'apprentissage, et les résultats obtenus montrent l'intérêt et l'efficacité de la méthode proposée.

Lors de la seconde partie, une méthode géométrique est proposée pour déterminer le point de vue le plus représentatif d'un objet en considérant son utilisation plutôt que son

esthétisme. Plus précisément, nous proposons d'évaluer la quantité de surfaces visibles ainsi que la saillance intrinsèque, tout en pondérant chaque sommet visible en fonction de l'angle de vue. Les résultats ont été validés par une étude utilisateur et utilisatrice, tout en assurant la cohérence avec la perception humaine.

Title : Analysis of 2D and 3D visual content : relevance evaluation of a 3D object point of view

Keywords : 3D models, depthmaps, 2D images, representation quality, 3D saliency, curvilinear saliency, relevance score, neural networks, confidence score, best view selection, user study.

Abstract :
In our three-dimensional world, objects are often represented in two dimensions, mainly visualized through 2D displays such as paper catalogs or computer screens. Selecting the viewpoint from which to observe a 3D object on a 2D display has a significant impact on its recognition, the identification of its characteristics and its functional use. This task is crucial in various applications related to video games, medical imaging, architecture and industrial design.

This thesis proposes to automatically select the most relevant 2D viewpoint for a given 3D object, by quantifying the relevance of a viewpoint as the visible « essential object characteristics », i.e. the information necessary to identify the object or its main attributes. We propose a method for automatically determining the most relevant 2D viewpoint by extracting and quantifying the essential information from each available viewpoint.

Two aspects were explored in this thesis. The first evaluates the relevance of fixed viewpoints offered by textured images relative to a 3D object. The geometry of this object is defined by a 3D model. The second aspect focuses on the analysis of a geometric 3D model defined by an untextured 3D mesh. We propose to automatically determine its most representative viewpoint, the one that allows unambiguous identification and understanding.

For the first axis of research, we are developing an approach based on the detection of salient object features while filtering out irrelevant information relative to appearance, such as texture, taking advantage of both 2D images and 3D models. We introduce a relevance score, derived from geometric attributes and photography-related recommendations, which enables us to rank all the viewpoints that are available. To validate the approach, we compare the relevance scores with confidence scores obtained from learning methods, and the results obtained show the interest and effectiveness of the proposed method.

In the second part, a geometric method is proposed to determine the most representative viewpoint of an object, considering its functionnal nature rather than its aestheticism. More specifically, we propose to evaluate the amount of visible surface as well as intrinsic saliency, while weighting each visible vertex according to viewing angle. The results were validated by a user study in order to guarantee the consistency with human perception. We carried out a comparison with two state-of-the-art methods for best view selection, based on mean curvature and saliency.

Remerciements

Je n'aurais certainement pas pu mener à bien ce projet de recherche, aussi enrichissant que satisfaisant, sans la présence et le soutien de nombreuses personnes. C'est pourquoi je tiens à exprimer ma gratitude envers tous ceux qui m'ont aidée dans cette aventure scientifique.

Pour commencer, je souhaite remercier Jean-Luc Mari d'avoir accepté d'être membre de mon jury de thèse et de le présider. Je souhaite également remercier Michèle Gouiffès et Romain Raffin pour avoir accepté d'être rapporteurs de ce manuscrit de thèse. J'ai particulièrement apprécié la qualité de leurs retours, de leurs commentaires constructifs, malgré le peu de temps disponible et la longueur de mon manuscrit, mais également des discussions que nous avons pu avoir lors de la soutenance.

Je souhaite remercier ma directrice de thèse, Géraldine Morin, et ma co-directrice de thèse Sylvie Chambon, qui sont bien plus que de simples encadrantes pour moi. Elles ont, tout au long de cette thèse, cru en moi et en mes capacités. Je souhaite particulièrement les remercier pour leur écoute et leur soutien.

Ensuite, je souhaite remercier tous les permanents et non permanents de l'équipe REVA, avec qui je partage de nombreux souvenirs, que ce soit dans les enseignements que nous avons pu donner ensemble, les multiples réunions et brainstorming, les repas, ou encore les parties de coinche et autres jeux de société. Je remercie particulièrement Thomas pour l'aide qu'il m'a apportée lors de la mise en place de ma soutenance de thèse et lors de la réalisation de mon étude utilisateur. Je remercie également Thierry de m'avoir appris à coder en javascript pour réaliser mon interface.

Plus personnellement, je voudrais remercier mes amis qui ont fait le déplacement pour venir me voir. Ceux de Toulouse, qui ont été formidables avec moi tout au long de ma thèse. Tous présents pour me changer les idées ou, plus joyeusement, fêter les grandes et les petites victoires. Des remerciements en particulier pour Travis, qui incarne l'optimisme et la positivité à toute épreuve. C'est un réel plaisir de travailler avec lui, et sa bonne humeur au quotidien. Je souhaite remercier une personne que je considère comme mon binôme, Rémy, qui m'a apporté au quotidien son soutien, son écoute, son aide et sa bonne humeur. C'était la personne la mieux placée pour me comprendre et m'épauler.

Je pense que je ne remercierai jamais assez mes parents de m'avoir toujours accompagnée et soutenue moralement, peu importe mes choix d'études. Bien qu'ils ne comprennent pas toujours ce que je fais, cela ne les a pas empêchés d'être toujours fiers de moi. Si j'ai pu réaliser cette thèse, c'est grâce, en partie, à mes merveilleux parents. Donc merci.

Enfin, j'aimerais terminer avec mon chéri, Valentin, qui m'a accompagnée et soutenue tous les jours, qui a cru en moi quand j'en avais besoin, qui m'a toujours écoutée sans

jamais se plaindre et qui m'a accompagnée par-delà le cercle polaire. Il a toujours su me motiver et prendre soin de moi, pour que je puisse être la plus efficace possible. La qualité de mes publications et de ma présentation n'aurait pas été la même sans lui. Donc merci pour tout, mon chéri.

Table des matières

I. Introduction générale 1

II. Méthodes d'évaluation de la pertinence en 2D et en 3D 9

1. Pertinence 2D . . . 11
   - 1.1. Qualité d'une image s'appuyant sur le confort visuel . . . 12
   - 1.2. Sélection de photographies dans une séquence d'images . . . 14
   - 1.3. Résumé de collection d'images . . . 15
     - 1.3.1. Résumé par classification . . . 16
     - 1.3.2. Résumé à l'aide d'apprentissage profond . . . 18
     - 1.3.3. Résumé adapté au cas des vidéos . . . 19
   - 1.4. Évaluation de la qualité esthétique d'image . . . 19
     - 1.4.1. Utilisation de caractéristiques bas niveau . . . 20
     - 1.4.2. Utilisation de caractéristiques haut niveau . . . 21
   - 1.5. Analyse des approches d'estimation de la pertinence en 2D . . . 23
2. Détection de points saillants en 2D et 3D . . . 24
   - 2.1. Approches basées premier ordre . . . 25
   - 2.2. Approches basées région . . . 25
   - 2.3. Approches basées second ordre . . . 26
   - 2.4. Notion d'analyse multi-échelle . . . 27
   - 2.5. Approches basées apprentissage . . . 29
   - 2.6. Points répétables en 2D et 3D . . . 32
   - 2.7. Bilan et choix . . . 35
3. Sélection de la meilleure vue en 3D . . . 37
   - 3.1. Surface . . . 39
   - 3.2. Silhouette . . . 40
   - 3.3. Profondeur . . . 41
   - 3.4. Notion de stabilité . . . 41
   - 3.5. Courbure et notion de saillance . . . 42
   - 3.6. Utilisation d'information *a priori* . . . 43
   - 3.7. Évaluation et combinaison des attributs géométriques . . . 44
   - 3.8. Approches basées apprentissage . . . 45
   - 3.9. Analyses des approches de sélection de la meilleure vue 3D . . . 47

4. Détection de saillance 3D . . . 48
4.1. Utilisation d'une caractérisation locale . . . 49
4.2. Utilisation d'une caractérisation globale . . . 50
4.3. Approches basées entropie . . . 51
4.4. Approches basées classification . . . 52
4.5. Approches basées apprentissage . . . 54
4.6. Bilan . . . 56

III. Points de vue restreints **61**

5. Pose favorable et image 2D la plus révélatrice d'un objet 3D . . . 63
5.1. Contexte . . . 63
5.2. Intérêt de la pose d'un objet 3D . . . 66
5.3. Attribut révélateur d'une image 2D . . . 70
5.4. Filtrage de la carte de saillance multi-échelle . . . 72
5.4.1. Notations utilisées pour les métriques de filtrage . . . 73
5.4.2. Utilisation d'une distance . . . 73
5.4.3. Statistiques classiques . . . 74
5.5. Classement d'images en fonction de la mise en valeur d'un objet 3D 75
5.5.1. Jeux de données 2D/3D . . . 75
5.5.2. Classement à partir des cartes de profondeur . . . 77
5.5.3. Classement des images révélatrices . . . 78
5.5.4. Limitations et perspectives . . . 80
6. Quantification améliorée de la pertinence . . . 85
6.1. Problématique . . . 85
6.2. Méthode déterministe proposée . . . 87
6.3. Choix des paramètres de la méthode déterministe . . . 88
6.4. Score de pertinence . . . 90
6.5. Méthodes utilisant un score de confiance classique en apprentissage profond . . . 92
6.5.1. Introduction aux réseaux de neurones . . . 92
6.5.2. Évaluation de l'esthétique d'une image . . . 94
6.5.3. Lien entre apprentissage humain et réseaux de neurones . . . 96
6.6. Construction des classements de référence . . . 98
6.6.1. Classements intra-dégradations . . . 100
6.6.2. Classements inter-dégradations . . . 102
6.7. Protocole d'évaluation . . . 103
6.8. Comparaisons quantitatives par type de dégradation . . . 105
6.8.1. Dégradation *augmentation* . . . 105
6.8.2. Dégradation *occultation* . . . 107

6.8.3. Dégradation *changement d'échelle* . . . 109
6.8.4. Dégradation *luminosité* . . . 111
6.8.5. Dégradation *flou gaussien* . . . 113
6.8.6. Bilan . . . 114
6.9. Comparaisons quantitatives sur la combinaison de dégradations . . . 115
6.10. Résultats qualitatifs . . . 118

IV. Points de vue non restreints **123**

7. Mesure de la pertinence d'une vue d'un objet 3D . . . 125
7.1. Introduction . . . 125
7.2. Extraction et combinaison d'attributs géométriques . . . 128
7.3. Méthodes de saillance 3D et formules d'angle testées . . . 131
7.3.1. Saillance intrinsèque 3D . . . 131
7.3.2. Détection des faces et sommets visibles . . . 132
7.3.3. Pondération de la saillance . . . 133
7.4. Proposition de validation . . . 134
7.4.1. Classification des techniques de validation . . . 134
7.4.2. Base de données . . . 134
7.4.3. Étude utilisateurs et utilisatrices . . . 134
7.4.4. Score de proximité . . . 137
7.5. Optimisation de notre approche . . . 140
7.5.1. Choix de la méthode de saillance 3D . . . 140
7.5.2. Étude d'ablation . . . 141
7.6. Comparaisons quantitatives et qualitatives . . . 141
8. Réalisation d'une étude utilisateurs et utilisatrices . . . 146
8.1. Contexte . . . 146
8.2. Présentation de l'interface . . . 149
8.2.1. Page d'introduction . . . 149
8.2.2. Définition du « bon » point de vue . . . 149
8.2.3. Tutoriel proposé et consignes . . . 150
8.2.4. Justification des choix des utilisateurs . . . 151
8.2.5. Stockage des données . . . 151
8.3. Fonctionnalités de l'interface . . . 151
8.4. Pré-traitement des caméras et des modèles 3D . . . 154
8.5. Post-traitement des données récoltées . . . 156
8.5.1. Réception des données . . . 156
8.5.2. Extraction d'informations statistiques . . . 158
8.5.3. Filtrage des participants . . . 159

8.5.4. Construction des histogrammes de popularité des points de vue . . . 161
8.6. Analyses et interprétations des histogrammes . . . 162
8.6.1. Cas général . . . 162
8.6.2. Vues accidentelles . . . 164
8.6.3. Vues occultées . . . 165
8.6.4. Objets avec des yeux . . . 165
8.6.5. Objets symétriques . . . 166
8.6.6. Objets non-familiers . . . 166
8.7. Comparaisons avec des images réelles . . . 168

V. Conclusion générale **173**

Résumés **203**
Popularized abstract . . . 203
Résumé vulgarisé . . . 203
Abstract . . . 205
Résumé . . . 205

Table des figures

I.1. Illustration de trois points de vue différents pour un même objet 3D. . . . 2
I.2. Deux contextes différents pour un même point de vue d'un objet 3D. . . . 3

II.1. Estimation de la qualité d'images en fonction du confort visuel. 12
II.2. Exemple de triplet d'images. 13
II.3. Exemples de séquences d'images. 14
II.4. Exemple illustrant une collection aléatoire de 32 images non triées. . . . 16
II.5. Exemple de graphes de scène. 17
II.6. Exemple de résumé obtenu par apprentissage profond. 18
II.7. Exemples de photographies classées selon leur qualité colorimétrique. . . . 21
II.8. Estimation de la qualité des images en fonction de leur contenu. 22
II.9. Illustration de scores de mémorabilité. 23
II.10. Illustration du détecteur Saddle. 27
II.11. Analyse multi-échelle dans une pyramide d'images (multi-résolution). . . . 28
II.12. Réseau convolutif pour la détection de points d'intérêt dans des images. . 30
II.13. Détection de points d'intérêt 2D/3D par saillance curviligne multi-échelle. 34
II.14. Comparaisons de deux points de vue pour deux opérations robotiques. . . 37
II.15. Illustrations de métriques pour déterminer la meilleure vue d'un squelette. 38
II.16. Évaluation de la qualité d'un point de vue 3D basé surface. 40
II.17. Classements de vues utilisant l'entropie des distributions des courbures. . 42
II.18. Pertinence d'un ensemble de vues en fonction de la saillance. 43
II.19. Illustrations de formes spatio-temporelles issues de séquences d'images. . . 45
II.20. Sélection de points de vue par étude statistique de collection d'images. . . 46
II.21. Sélection de points de vue utilisant des croquis humains. 47
II.22. Caractérisation en s'appuyant sur les sommets, les extrémités et les patches. 50
II.23. Visualisation de l'angle tropical utilisé pour la saillance 3D. 51
II.24. Estimation de la saillance 3D à l'aide de champs aléatoires conditionnels. . 52
II.25. Étapes d'une méthode de détection de saillance 3D basée classification. . . 53
II.26. Illustrations du concept de saillance tactile. 55
II.27. Schéma et résultats d'une expérience de suivi du regard. 56
II.28. Approche validée par des cartes de fixation du regard. 57

III.1. Illustrations de la notion de pose favorable. 65
III.2. Illustrations de la notion d'image révélatrice. 65
III.3. Deux critères pour évaluer la qualité de la pose d'un objet 3D. 67
III.4. Chaîne de traitement pour classer les orientations favorables d'un objet 3D. 68
III.5. Illustration de différentes poses. 69
III.6. Chaîne de traitement pour classer les positions favorables d'un objet 3D. . 69
III.7. Influence de la résolution sur le nombre de points saillants disponibles. . . 70
III.8. Chaîne de traitement pour classer les images selon une propriété révélatrice. 71
III.9. Saillance curviligne pour une image et sa carte de profondeur associée. . . 72
III.10. Mise en correspondance 2D/3D où le modèle 3D diffère de celui en 2D. . 75
III.11. Exemples d'images et de modèles 3D de la base de données Pix3D. 76
III.12. Classements des vues d'un même objet en fonction de son orientation. . . 77
III.13. Classements des vues d'un même objet en fonction de sa position. 78
III.14. Classement des vues d'un même objet en fonction de sa pose. 78
III.15. Classements des images révélatrices (suivant deux distances). 81
III.16. Classements des images révélatrices (suivant des statistiques classiques). . 82
III.17. Illustrations des limites de l'approche déterministe proposée. 83
III.18. Classement d'images réelles en fonction de leur pertinence. 85
III.19. Chaîne de traitement pour classer des images par pertinence. 87
III.20. Processus d'estimation des scores de pertinence. 89
III.21. Différences entre un point de vue pertinent et non pertinent. 90
III.22. Scores de confiance issu de l'apprentissage profond. 96
III.23. Illustration du concept de typicité. 97
III.24. Illustrations des cinq dégradations considérées. 98
III.25. Classement de référence avec uniquement des *augmentations*. 100
III.26. Classement de référence avec uniquement des *occultations*. 100
III.27. Classement de référence avec uniquement des *changements d'échelle*. . . . 101
III.28. Classement de référence avec uniquement des changements de *luminosité*. 101
III.29. Classement de référence avec uniquement des applications de *flou gaussien*. 102
III.30. Évaluation des méthodes de classement. 104
III.31. Résultats associés à la dégradation *augmentation*. 106
III.32. Résultats associés à la dégradation *occultation*. 108
III.33. Résultats associés à la dégradation *changement d'échelle*. 110
III.34. Résultats associés à la dégradation *luminosité*. 112
III.35. Résultats associés à la dégradation *flou gaussien*. 113
III.36. Résultats associés à la combinaison de dégradations. 116
III.37. Comparaisons visuelles entre le score de pertinence et le score de confiance. 120
III.38. Classements d'images réelles obtenus avec l'approche déterministe. 121

IV.1. Illustration du biais lié à l'esthétisme avec l'exemple d'une tasse. 127
IV.2. Illustration du biais de variabilité limitée des points de vue accessibles. . . 127
IV.3. Chaîne de traitement pour sélectionner la meilleure vue d'un objet 3D. . . 129
IV.4. Sélection des faces visibles selon un point de vue. 132
IV.5. Sélection des sommets visibles selon un point de vue. 133
IV.6. Exemple de modèles 3D de la base de maillages 3D utilisée. 136
IV.7. Répartition des points de vue étudiés pour un objet 3D. 136
IV.8. Histogramme des préférences humaines pour les points de vue étudiés. . . 137
IV.9. Zone d'influence du score de proximité. 138
IV.10. Illustration des vues particulières considérées comme pertinentes. 139
IV.11. Résultats quantitatifs de l'étape d'optimisation. 140
IV.12. Étude d'ablation. 141
IV.13. Comparaisons quantitatives des méthodes étudiées. 142
IV.14. Comparaisons qualitatives des méthodes étudiées. 144
IV.15. Pages d'introduction de l'interface utilisateurs. 150
IV.16. Page principale de l'interface utilisateurs. 151
IV.17. Pages du tutoriel de l'interface utilisateurs. 153
IV.18. Base de données et répartition des caméras proposée aux utilisateurs. . . 154
IV.19. Sauvegarde des données. 156
IV.20. Extrait des interactions enregistrées dans un fichier JSON d'un utilisateur. 157
IV.21. Répartition de modèles 3D traités et analysés durant l'étude. 158
IV.22. Données statistiques relatives aux 203 participants. 160
IV.23. Influence du choix des caméras. 161
IV.24. Histogramme de popularité dans le cas général. 163
IV.25. Histogramme de popularité pour les vues accidentelles. 164
IV.26. Histogramme de popularité pour les vues occultées. 165
IV.27. Histogramme de popularité pour les objets avec des yeux. 166
IV.28. Histogramme de popularité pour les objets symétriques. 167
IV.29. Histogramme de popularité pour les objets non familiers. 167
IV.30. Influence des conditions réelles sur les préférences des utilisateurs. 169
IV.31. Exemple d'un objet avec et sans texture. 170

V.1. Modèle 3D texturé. 175
V.2. Illustration du concept de Compréhension de scène. 176
V.3. Illustrations de scènes dynamiques de bandes dessinées. 177

Liste des tableaux

II.1. Publications sur la qualité esthétique d'images (bas niveau). 20
II.2. Publications sur la qualité esthétique d'images (haut niveau). 21

III.1. Paramètres de création des classements de référence avec des dégradations mixtes. 103
III.2. Coefficients de corrélation associés à la dégradation *augmentation*. 107
III.3. Coefficients de corrélation associés à la dégradation *occultation*. 109
III.4. Coefficients de corrélation associés à la dégradation *changement d'échelle*. 111
III.5. Coefficients de corrélation associés à la dégradation *luminosité*. 112
III.6. Coefficients de corrélation associés à la dégradation *flou gaussien*. 114
III.7. Bilan des performances de chaque méthode par rapport aux différents types de dégradations étudiés. 115
III.8. Coefficients de corrélation associés à la combinaison de dégradations. . . . 117

IV.1. Techniques utilisées pour la validation des méthodes de sélection de la meilleure vue d'objet 3D. 135
IV.2. Détails des études utilisateurs de l'état de l'art. 147
IV.3. Récapitulatif des fonctionnalités disponibles. 152
IV.4. Étiquettes des caméras en fonction de leur position. 155

Chapitre I

Introduction générale

Bien que la plupart des objets que nous manipulons dans notre monde physique soient en trois dimensions (3D), lorsque nous les représentons, ces objets sont visibles en deux dimensions (2D). En effet, les objets 3D avec lesquels nous souhaitons interagir sont principalement visualisés à travers des supports 2D, impressions en papier ou écrans divers. Par conséquent, il n'est pas possible de visionner en 2D un objet 3D dans son intégralité. La représentation en 2D d'un objet 3D implique alors la nécessité de choisir un point de vue spécifique.

Ce point de vue va varier en fonction du contexte d'application. Ces points de vue ont une influence sur la quantité d'information et de caractéristiques visibles. Effectivement, lorsque la vue proposée d'un objet 3D contient des auto-occultations, rendant ainsi des parties non visibles, cf. Figure I.1b ou encore des potentielles superpositions de parties, créant alors une ambiguïté sur la perception de la profondeur cf. Figure I.1c, la quantité d'éléments visibles peut fortement varier. Il est donc essentiel de bien choisir son angle de vue en fonction de la tâche à réaliser.

Par ailleurs, cette quantité d'information disponible peut également varier même si le point de vue est fixé : cela dépend de l'environnement dans lequel l'objet est placé. Si nous considérons les deux images de la Figure I.2, le point de vue offert du canapé est identique, mais la quantité d'éléments visibles diffère. Dans l'image de droite, l'objet est placé dans un environnement avec d'autres objets qui l'occultent partiellement, et donc limitent l'accès à son information caractéristique.

▷ Critères de sélection possibles :

Le choix d'un point de vue au sein d'un support 2D pour un objet 3D peut être soumis à divers critères en fonction du contexte dans lequel il est réalisé. Dans la littérature, de nombreuses approches ont été développées pour la sélection de points de vue optimaux, que ce soit pour des images [Nishiyama 11, Xu 20] ou des modèles 3D [Secord 11, Nouri 15].

De manière générale, dans la littérature, nous pouvons trouver des méthodes de sélection de vues 2D qui s'appuient sur :

- Un *critère esthétique*, comme dans [Chang 16] : l'image la plus agréable à regarder esthétiquement. Par exemple, à la suite d'un voyage, nous pourrions souhaiter ne garder que les clichés les plus esthétiquement plaisants (cadrage avantageux, couleurs

(a) (b) (c)

FIGURE I.1. – Illustration de trois points de vue différents pour un même objet 3D. Ces trois points de vue présentent des détails différents et complémentaires : En (a) tous les éléments caractéristiques d'une tasse sont visibles à savoir, l'anse et le relief de la tasse, alors qu'en (b) l'anse de la tasse n'est pas visible, il s'agit d'une auto-occultation, mais nous pouvons observer l'autre côté de la tasse et enfin, en (c) nous pouvons voir l'anse de face, mais nous considérons ce point de vue d'accidentel. En effet, il y a une ambiguïté sur la perception de la profondeur. Dans le cadre de cette thèse, ce que nous souhaitons, c'est mettre en avant le premier point de vue car c'est celui qui contient le plus d'information pour reconnaître l'objet et comprendre comment l'utiliser.

harmonieuses, luminosité plaisante) en éliminant toutes les autres photographies. La plupart de ces méthodes se concentrent sur des critères esthétiques afin d'imiter les préférences subjectives des humains.

- Un *critère de compression*, comme dans [Ma 19] : l'image la plus légère d'un point de vue de mémoire. L'objectif peut être de stocker la représentation 2D en prenant le moins de place mémoire possible.
- Un *critère de confort visuel*, comme dans [Ben Amor 10] : nous souhaitons éliminer les images de mauvaise qualité, celles qui sont floues, peu contrastées ou peu nettes à cause de la présence de bruit.
- Un *critère de pertinence*, qui correspond à ce qui nous intéresse dans ce travail de thèse : avoir accès à une quantité importante d'information caractéristique visible de l'objet. Cette information permet d'identifier et d'avoir une compréhension globale de l'objet sans ambiguïté.

Les approches esthétiques ou celles qui favorisent le confort visuel traitent souvent les images dans leur globalité, en s'appuyant sur des attributs liés à la couleur, sans prendre en compte spécifiquement les caractéristiques propres de l'objet d'intérêt. De plus, ces méthodes sont efficaces sur des images corrélées, c'est-à-dire des images d'une même scène, mais avec un angle de vue légèrement différent. Notre étude se concentre spécifiquement sur le critère de pertinence, visant à quantifier la mise en valeur d'une image relative à

(a) Vue 2D sans occultation (b) Vue 2D avec occultations

FIGURE I.2. – Illustration de deux contextes différents pour un même point de vue d'un objet 3D. La quantité de détails visibles est différente en raison de la présence d'objets supplémentaires occultants.

un objet d'intérêt 3D spécifique, introduisant ainsi un critère différent et plus local à la sélection d'images.

▷ Problématique - Définition de la pertinence d'une vue 2D :

Objectif

À partir d'un objet 3D, être capable de quantifier la pertinence, et non l'esthétisme, offert par une visualisation 2D d'un objet étudié.

Notre objectif est de concevoir une méthode capable de déterminer automatiquement la vue 2D la plus pertinente pour un objet 3D, c'est-à-dire celle qui représente au mieux l'objet 3D, mettant en évidence ses parties les plus informatives. Cette vue permet une identification sans ambiguïté de l'objet, et une compréhension de son fonctionnement, le cas échéant. Cette pertinence est ce que nous avons défini comme étant l'« *information essentielle* ». Précisément, l'information essentielle d'une vue ou d'un objet constitue le plus petit ensemble englobant les éléments caractéristiques et fondamentaux nécessaires à la définition et à la compréhension globale de l'objet concerné. Cette information extraite est intrinsèquement plus concise que la totalité de l'information initiale associée à l'objet. De plus, il est important de rappeler que notre problématique ne s'attache pas à l'esthétisme des représentations visuelles, ni aux préférences subjectives des individus. Au contraire, nous souhaitons nous concentrer sur la pertinence et la représentativité d'une vue 2D d'un objet 3D, en s'appuyant sur des caractéristiques objectives disponibles dans les visuels 2D, tels que les attributs géométriques de notre objet. Ce critère de pertinence nous permet, en résultat final, de classer automatiquement les différentes vues.

Enfin, nous souhaitons que les vues 2D sélectionnées garantissent une reconnaissance sans ambiguïté par les êtres humains. Plus précisément, les vues sélectionnées doivent refléter celles qu'un être humain aurait choisies en fonction de ses besoins spécifiques pour une tâche donnée. En d'autres termes, cela signifie que nous souhaitons qu'elles soient porteuses de sens pour les utilisateurs et les utilisatrices.

▷ Complémentarité des données étudiées en 2D et en 3D :

L'objectif principal de notre étude est de mesurer la pertinence d'une vue 2D relative à un objet 3D, en se focalisant sur les deux représentations les plus communes et évidentes d'un objet 3D : les images de cet objet et le modèle 3D de l'objet lui-même. Nous travaillons avec un ensemble d'images très différentes que ce soit d'un point de vue des textures, des lieux, des objets environnants, des éclairages, etc. L'objectif est de mesurer à quel point une image représente correctement l'objet.

En 3D, nous avons fait le choix de manipuler des maillages, sans texture, sans aucun autre objet visible dans la scène, et nous supposons que les utilisateurs et utilisatrices peuvent les voir sur un écran 2D. Ainsi, en 3D, il n'y a ni risque d'occultation par d'autres objets ni variations de lumière ou de couleur.

Une différence entre ces deux modalités réside dans leur capacité à offrir une variété de points de vue différents. En utilisant des images, le nombre de vues est limité par le nombre initial d'images disponibles et la sélection du meilleur point de vue est donc fortement dépendant de la pertinence des points de vue des images acquises. Contrairement aux images 2D, lorsque nous travaillons directement avec le modèle 3D, l'ensemble des vues disponibles est potentiellement infini. En effet, l'objet 3D peut être librement tourné dans toutes les directions, offrant ainsi une variété illimitée de points de vue.

▷ Intérêts applicatifs de la sélection de vues pertinentes en 2D et en 3D :

Un point de vue judicieusement sélectionné facilite la compréhension et l'identification rapide de l'objet, offrant aux personnes utilisatrices un sentiment de satisfaction et de confort. Cette rapidité dans la reconnaissance est d'autant plus importante dans des applications commerciales, où elle peut se traduire par des économies de temps et d'argent, voire la multiplication de projets réalisables plus rapidement.

Par exemple, dans le domaine des jeux vidéos ou de la réalité virtuelle, un choix de point de vue optimal pour les objets interactifs contribue au confort et au plaisir des joueurs et des joueuses, suscitant ainsi l'envie de continuer ou d'explorer d'autres opus. En revanche, un point de vue aléatoire peut entraver la reconnaissance de l'objet. Il complexifie également la compréhension de l'utilité de l'objet, générant ainsi de la frustration et demandant à la personne utilisatrice de consacrer davantage de temps et d'efforts pour saisir pleinement les fonctionnalités de l'objet.

Dans le domaine médical, la réalité augmentée peut tirer profit de l'évaluation automa-

tique des meilleures vues d'organes pour une visualisation précise, aidant ainsi les médecins lors des diagnostics puis les chirurgiens et chirurgiennes lors d'interventions.

Les architectes et, plus généralement, les concepteurs et les conceptrices peuvent également trouver un avantage dans notre approche, en intégrant des méthodes d'évaluation de la pertinence des vues dans les logiciels de modélisation architecturale pour une représentation optimale des bâtiments et des structures.

Dans le secteur industriel, les ingénieurs et ingénieures peuvent optimiser la visualisation de prototypes 3D grâce à l'évaluation de la pertinence des vues, améliorant ainsi la communication au sein des équipes de conception et avec les personnes clientes.

Tous les exemples que nous venons de fournir concernent des applications manipulant directement le modèle 3D de l'objet étudié. Dans certains domaines, ce sont directement des images 2D qui sont utilisées.

En effet, dans le secteur de la sécurité et de la surveillance, une sélection précise d'images serait pertinente pour la reconnaissance faciale, que ce soit pour l'identification des individus ou la recherche de personnes suspectes. Une pré-sélection des points de vue pertinents relatifs à une personne pourrait garantir une acquisition optimale des images contenant les visages spécifiques nécessaires.

De même, dans le domaine médical, tout type d'imagerie médicale bénéficierait de la sélection d'images montrant des régions anatomiques spécifiques sans occultations et facilitant ainsi le diagnostic et l'étude de l'évolution des pathologies.

Enfin, dans le domaine publicitaire, la création de supports visuels attrayants repose sur la sélection d'images qui mettent en valeur les caractéristiques distinctives des produits, contribuant ainsi à capter l'attention des consommateurs et des consommatrices.

Ces considérations soulignent l'importance des problématiques posées dans cette thèse dans des applications et des contextes variés, où l'évaluation automatique de la pertinence des représentations 2D contribue directement à l'efficacité, à la satisfaction des utilisateurs et des utilisatrices ainsi qu'à la réussite des projets.

Comme évoqué précédemment, cette thèse se focalise sur deux axes d'étude visant à évaluer la pertinence d'une vue 2D d'un objet 3D. Ces axes présentent des méthodes similaires qui reposent sur l'extraction de l'information essentielle liée à l'objet 3D, mais à partir de données différentes : soit à partir d'images 2D de l'objet en utilisant son modèle 3D correspondant, soit à partir de son maillage 3D directement.

▷ Quantification de la pertinence d'une vue 2D en s'appuyant sur des images et un modèle 3D de l'objet représenté :

Nous rappelons que nous cherchons à extraire l'information essentielle, c'est-à-dire tout ce qui doit être conservé pour pouvoir comprendre et identifier correctement un objet donné aussi bien en 2D, dans chacune des vues images étudiées mais également en 3D,

dans les différentes vues du modèle, et plus exactement, les différentes cartes de profondeur que nous avons décidé d'exploiter. Que ce soit en 2D ou en 3D, nous avons fait le choix de récupérer cette information essentielle par extraction de points saillants à l'aide d'un détecteur basé sur la saillance curviligne [Rashwan 19] [1]. La sélection de ce détecteur est lié au fait qu'il est robuste, c'est-à-dire répétable d'une modalité à une autre, dans des conditions difficiles par exemple, dans le cas d'objets occultants. De plus, l'intérêt du détecteur utilisé réside dans le fait que nous pouvons nous concentrer sur la géométrie de l'objet, élément qui nous semble le plus important pour caractériser la fonctionnalité ainsi que l'identité d'un objet. Nous sommes alors capables de distinguer les caractéristiques géométriques essentielles de l'objet étudié des autres caractéristiques non essentielles liées à la texture d'autres objets. L'originalité de la méthode proposée réside dans l'utilisation de ce résultat pour calculer un « *score de pertinence* » indépendant des couleurs et des textures, correspondant à une combinaison entre l'information de saillance propre à l'objet et des aspects liés à la photographie tels que la place occupée par l'objet et sa taille par rapport au reste de l'image. Pour valider nos résultats, nous avons élaboré un protocole original impliquant la dégradation progressive des images avec des altérations visuelles réalistes. Plus une image est dégradée, moins elle est pertinente vis-à-vis de l'objet étudié. L'objectif est de démontrer la capacité de notre méthode à retrouver l'ordre des dégradations. Pour évaluer nos performances, nous avons comparé notre *score de pertinence*, avec le *score de confiance* généré par des méthodes d'apprentissage profond. Cette comparaison a été inspirée de précédents travaux dans la littérature, comme ceux de [Lake 15], qui ont déjà mis en compétition les performances humaines et celles des réseaux neuronaux. Cette comparaison se justifie par l'idée que la mise en valeur d'un objet dans une image correspond à un *score de confiance* élevé attribué par un classifieur.

▷ Quantification de la pertinence d'une vue 2D directement à partir du modèle 3D :

La dernière partie des travaux de cette thèse s'intéresse au cas où seul le modèle 3D est utilisé pour déterminer la vue la plus pertinente. Cela signifie que nous ne travaillons plus avec un ensemble de vues restreint. Le domaine de recherche le plus proche concerne ce que l'on appelle la sélection de la meilleure vue ou *Best View Selection*. Comme en 2D, les méthodes existantes cherchent à imiter les choix esthétiques humains. De plus, les approches proposées sont rarement évaluées de manière quantitatives. Notre travail se distingue en cherchant, d'une part, à considérer l'information essentielle, et d'autre part, en proposant une validation la plus rigoureuse possible. Ainsi, pour évaluer la pertinence d'un point de vue directement à partir du maillage 3D, nous avons développé une méthode combinant la saillance intrinsèque des objets avec des attributs géométriques spécifiques à chaque point de vue tels que la quantité de surface visible. Pour chaque vue, la saillance intrinsèque des

1. Il s'agit d'un détecteur proposé et étudié au sein de l'équipe REVA où cette thèse a été réalisée.

sommets visibles est pondérée par leur angle de vue par rapport à la position de la caméra. Ceci nous permet de favoriser les éléments visibles les plus pertinents dans chaque vue, formant ainsi l'information essentielle disponible dans la vue. Notre quantification de la mise en valeur des objets 3D repose uniquement sur des attributs géométriques. En l'absence de bases de données adaptées pour valider ce travail, nous avons réalisé une étude utilisateur et utilisatrice. Cette dernière nous a permis de former notre vérité terrain en récoltant les vues d'objets 3D considérées comme étant les plus pertinentes selon les personnes utilisatrices.

▷ Organisation du manuscrit :

Ainsi, cette thèse va permettre le développement suivant : un premier chapitre introductif de l'état de l'art, c'est-à-dire des méthodes d'évaluation de la pertinence en 2D et en 3D, un second chapitre qui correspond à ce que nous avons détaillé dans le paragraphe quantification de la pertinence d'une vue 2D en s'appuyant sur des images et un modèle 3D de l'objet représenté de cette introduction, un troisième chapitre contenant ce qui est décrit dans le paragraphe quantification de la pertinence d'une vue 2D directement à partir du modèle 3D avant de conclure dans le dernier chapitre.

Chapitre II

Méthodes d'évaluation de la pertinence en 2D et en 3D

Sommaire

1. Pertinence 2D . . . 11
   - 1.1. Qualité d'une image s'appuyant sur le confort visuel . . . 12
   - 1.2. Sélection de photographies dans une séquence d'images . . . 14
   - 1.3. Résumé de collection d'images . . . 15
     - 1.3.1. Résumé par classification . . . 16
     - 1.3.2. Résumé à l'aide d'apprentissage profond . . . 18
     - 1.3.3. Résumé adapté au cas des vidéos . . . 19
   - 1.4. Évaluation de la qualité esthétique d'image . . . 19
     - 1.4.1. Utilisation de caractéristiques bas niveau . . . 20
     - 1.4.2. Utilisation de caractéristiques haut niveau . . . 21
   - 1.5. Analyse des approches d'estimation de la pertinence en 2D . . . 23
2. Détection de points saillants en 2D et 3D . . . 24
   - 2.1. Approches basées premier ordre . . . 25
   - 2.2. Approches basées région . . . 25
   - 2.3. Approches basées second ordre . . . 26
   - 2.4. Notion d'analyse multi-échelle . . . 27
   - 2.5. Approches basées apprentissage . . . 29
   - 2.6. Points répétables en 2D et 3D . . . 32
   - 2.7. Bilan et choix . . . 35
3. Sélection de la meilleure vue en 3D . . . 37
   - 3.1. Surface . . . 39
   - 3.2. Silhouette . . . 40
   - 3.3. Profondeur . . . 41
   - 3.4. Notion de stabilité . . . 41
   - 3.5. Courbure et notion de saillance . . . 42

3.6. Utilisation d'information *a priori* 43
3.7. Évaluation et combinaison des attributs géométriques 44
3.8. Approches basées apprentissage 45
3.9. Analyses des approches de sélection de la meilleure vue 3D 47
4. Détection de saillance 3D 48
4.1. Utilisation d'une caractérisation locale 49
4.2. Utilisation d'une caractérisation globale 50
4.3. Approches basées entropie 51
4.4. Approches basées classification 52
4.5. Approches basées apprentissage 54
4.6. Bilan 56

Pourquoi la sélection des meilleurs points de vue est-elle importante ?

Un grand nombre d'objets en 3D sont utilisés quotidiennement dans divers domaines tels que le développement de jeux vidéo, la conception assistée par ordinateur ou encore la décoration d'intérieur, pour ne citer que quelques exemples. L'utilisation de ces modèles 3D repose souvent sur l'exploration de grandes bases de données et également sur la manipulation manuelle du modèle afin de l'utiliser avantageusement. Cette tâche est fastidieuse. Dans ce contexte, la sélection automatisée de meilleurs points de vue doit permettre de diminuer le temps de manipulation et également de reconnaître ou de comprendre plus facilement le modèle 3D.

Nous rappelons que notre objectif principal est de concevoir une méthode capable de déterminer automatiquement la vue 2D la plus pertinente pour un objet 3D, en se concentrant sur les deux représentations les plus courantes : les images de l'objet et le modèle 3D de celui-ci. Identifier la vue 2D la plus pertinente revient à choisir la perspective qui offre la meilleure représentation de l'objet, mettant en évidence ses caractéristiques essentielles et facilitant ainsi la prise de décision ou l'analyse.

Ainsi, dans ce chapitre, nous présentons les méthodes existantes en sélection de meilleurs points de vue 2D d'un objet 3D. Nous examinons en détails les différentes approches proposées dans la littérature, en mettant particulièrement l'accent sur leur capacité à mesurer ce qui nous intéresse le plus, à savoir la pertinence.

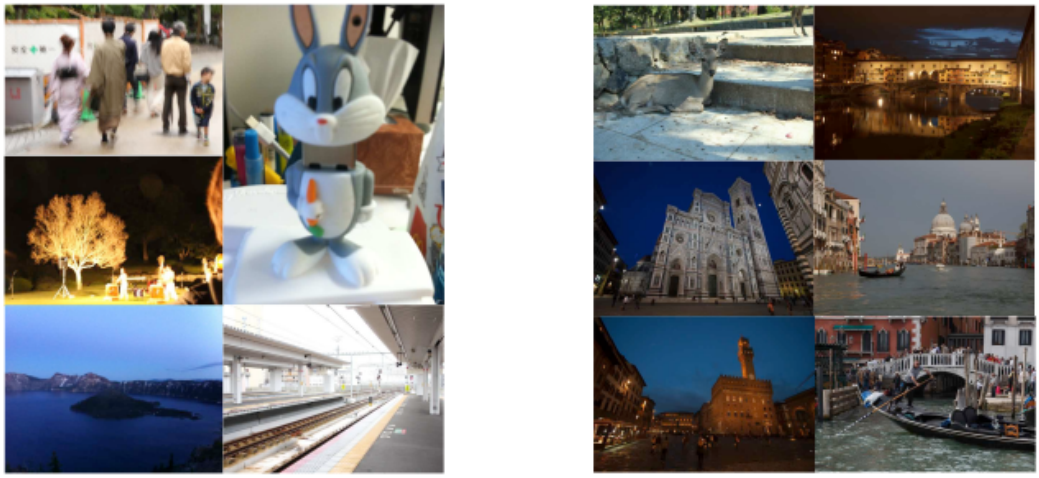
Ce chapitre sera structuré en quatresections. Dans la section 1, nous aborderons la pertinence des vues 2D issues d'images. Ensuite, nous explorerons les méthodes de détection de points saillants en 2D et en 3D dans la section 2, en reprenant la classification proposée dans [Chambon 20] et en la complétant avec les méthodes s'appuyant sur un apprentissage. La section 3, se concentrera sur la sélection de la meilleure vue et nous passerons en revue les diverses modalités et techniques utilisées, ainsi que les approches faisant appel à l'apprentissage et l'extraction d'attributs géométriques. Enfin, dans la section 4, nous présenterons les méthodes de détection de la saillance 3D existantes.

1. Pertinence 2D

Il existe différentes manières d'évaluer la qualité d'une image. Dans notre étude, nous souhaitons estimer la qualité d'une image en s'appuyant sur sa pertinence par rapport à l'objet qu'elle représente et en utilisant le maillage 3D de celui-ci. Cette tâche est une problématique rarement abordée dans la littérature. En effet, les travaux existants se concentrent généralement sur l'évaluation globale de la qualité d'une image, par exemple, dans le cas de la restauration d'images [Chambah 03, Chambah 07]. Ainsi, de nombreuses

mesures ont été développées pour reproduire au mieux le jugement humain et se rapprocher de la perception visuelle humaine. Cette qualité perçue peut être considérée selon plusieurs objectifs divers que nous allons aborder dans les sections suivantes : le confort visuel, la sélection de série de photographies, et plus généralement le tri automatique d'images et enfin le résumé d'images. La dernière section nous permettra de proposer une synthèse de cet état de l'art en 2D.

1.1. Qualité d'une image s'appuyant sur le confort visuel



(a) Images de mauvaise qualité.

(b) Images de bonne qualité.

FIGURE II.1. – Estimation de la qualité d'images en fonction du confort visuel [Lu 15]. Nous reprenons les illustrations fournies dans cet article. Le réseau de neurones entraîné dans cette publication permet de distinguer les images de faible qualité comme c'est le cas pour l'ensemble d'images présenté en (a), caractérisées par la présence de flou, un contraste faible ou une surexposition des images de haute qualité, présentées en (b).

Dans le cadre de notre étude sur l'évaluation de la qualité d'une image en fonction de l'objet qu'elle représente, nous examinons également les travaux qui évaluent le confort visuel, englobant divers aspects techniques des images tels que le flou, le contraste ou la netteté. La qualité qui est ici estimée, peut être considérée comme une qualité dite fonctionnelle, correspondant à l'estimation du confort visuel associé à une image. Ces attributs étudiés peuvent considérablement influencer la qualité perçue d'une image, affectant ainsi sa compréhension et réduisant la mise en valeur de certains objets qui y sont présents. Dans la littérature, de nombreuses méthodes ont été développées pour évaluer le confort visuel d'une image, cherchant à reproduire au mieux la perception humaine. Par exemple, certaines approches classiques consistent à évaluer la qualité de l'image en mesurant le niveau de flou ou la présence de bruit, comme dans [Ben Amor 10]. Dans ces travaux, une

nouvelle métrique objective est proposée. Elle reflète au mieux le jugement d'un observateur ou d'une observatrice, c'est-à-dire une métrique qui permet de reproduire sa sensibilité au contraste, au mouvement, au flou et aux couleurs. Dans les travaux de [Ouni 09], une mesure de perception de la qualité est construite en s'appuyant sur l'étude des écarts colorimétriques entre une image originale et une version modifiée, restaurée, de cette image. Contrairement à de précédents travaux comme ceux de [Luo 00] qui ont proposé des métriques de différences de couleur s'appuyant sur des différences locales pixel à pixel, la métrique proposée traite les images dans leur intégralité. Les scores obtenus permettent d'estimer l'efficacité de la restauration appliquée.

FIGURE II.2. – Exemple de triplet d'images (une de référence, au milieu et deux dégradées, à gauche et à droite) utilisé dans [Prashnani 18]. Dans leur étude, pour ce triplet, 88% des utilisateurs et utilisatrices n'ont pas eu de difficulté à identifier l'image à droite comme étant la plus proche perceptuellement de l'image de référence.

Lorsqu'il est possible de réaliser un apprentissage, nous pouvons citer le travail de [Lu 15] qui propose d'évaluer ce que nous avons nommé la qualité fonctionnelle. En effet, ils ont collecté un ensemble de photographies étiquetées manuellement en fonction de leur niveau de qualité fonctionnelle. Les photographies de faible qualité fonctionnelle correspondent aux images avec une surexposition ou une sous-exposition, un faible contraste de couleur ou encore avec un flou ou un bruit important, comme illustre dans la Figure II.1. D'autres travaux estiment la qualité fonctionnelle d'une image par comparaison [Prashnani 18]. L'originalité de cette méthode est d'effectuer l'apprentissage à partir de paires d'images dégradées. Ainsi, au lieu d'assigner un score de qualité à chaque image dégradée et de les apprendre individuellement, ce travail s'appuie sur des triplets constitués d'une image de référence et de deux versions dégradées, cf. l'exemple de la Figure II.2. Lors d'une étude avec des utilisateurs et des utilisatrices, les personnes ont été invitées à choisir, parmi les deux versions dégradées de chaque triplet, celle qui était perceptuellement la plus proche de l'image de référence. Cette tâche fait donc l'hypothèse qu'il est beaucoup plus facile de comparer deux images données et d'identifier celle qui est la plus similaire à une référence que d'attribuer des scores de qualité à chacune. Bien que la comparaison

par paires soit de plus en plus utilisée en raison de sa simplicité et de sa fiabilité, le nombre croissant de paires à évaluer représente un défi. Pour résoudre ce problème, des méthodes d'échantillonnage, telle que celle de [Mohammadi 23], *Predictive Sampling for Pairwise Comparison*, PS-PC, ont été développées. L'objectif étant de demander moins d'annotations de paires d'images à un expert ou une experte, le principe est de sélectionner automatiquement un sous-ensemble de paires à annoter, et d'estimer les annotations pour le reste des paires non annotées. Pour cela, un classifieur est entraîné pour identifier les paires les plus pertinentes à être évaluées par des utilisatrices ou des utilisateurs, c'est-à-dire, celles dont la distinction entre les deux est la plus visible pour un être humain.

1.2. Sélection de photographies dans une séquence d'images

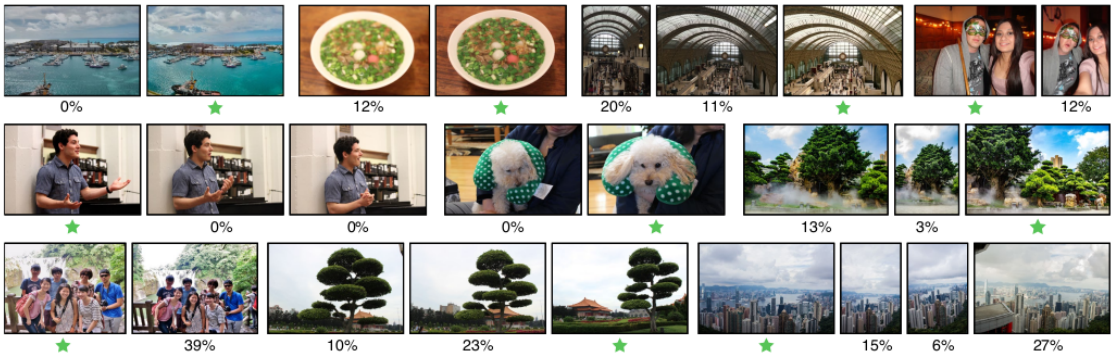


FIGURE II.3. – Exemples de séquences d'images extraites de [Chang 16] : pour chaque séquence, la photographie avec une étoile est celle préférée par la majorité des utilisateurs et utilisatrices, tandis que le pourcentage sous chaque photographie indique la fraction de personnes qui préfère cette photographie à la photographie étoilée de la même séquence.

Le domaine de la *Series photo selection* ou sélection de photographies représentatives au sein de séquences d'images vise à sélectionner et conserver l'image ou les images de meilleure qualité parmi une série de photographies, autrement dit une collection d'images corrélées donnée. Ces images sont corrélées car elles se rapportent au même sujet photographié mais sous différents angles, éclairages ou contexte, comme dans une vidéo ou une séquence d'images, comme illustré dans la Figure II.3. Cette tâche, laborieuse et pénible, peut donc être facilitée en utilisant des outils automatiques de sélection. En effet, les auteurs et autrices de [Drucker 03] ont élaboré une interface utilisateur et utilisatrices visant à améliorer l'expérience en offrant une meilleure visualisation des photographies et à simplifier les interactions lors du tri des photographies. De même, dans [Jacobs 10], les auteurs ont introduit le concept de *cosaliency* ou co-saillance s'appuyant sur les différences bas niveau, telles que le bruit, la mise au point, le flou ou encore l'exposition, entre des paires d'images. Par ailleurs, les auteurs de [Zhu 14] se sont concentrés sur la sélection de

portraits les plus attrayants parmi une vaste collection de photographies en exploitant les caractéristiques du visage. Plus précisément, ils extraient les caractéristiques du visage en utilisant le descripteur HOG [Dalal 05], *Histogram of Oriented Gradients* ou Histogramme des gradients orientés, qui permet une analyse multi-échelle des propriétés visuelles des expressions faciales dans différentes parties du visage à différentes échelles. Plus précisément, ils définissent cinq boîtes englobantes qui correspondent aux deux yeux, aux sourcils et aux rides d'expressions autour des sourcils, la bouche et l'ensemble du visage.

Avec l'essor des techniques par apprentissage profond dans les années 2010, les approches de sélection d'images ont également bénéficié de cette avancée. La méthode présentée dans [Chang 16] utilise un réseau convolutif avec une architecture siamoise : les caractéristiques sont d'abord extraites de chaque paire d'images par deux réseaux avec des poids partagés. Ensuite, la différence entre les caractéristiques est transmise à la partie commune du réseau qui détermine en sortie laquelle des deux images est préférée. Dans leurs travaux [Kuzovkin 17, Kuzovkin 18], les auteurs et autrices proposent un nouveau processus qui facilite la sélection des photographies en évaluant la qualité de l'image tout en tenant compte du contexte de l'image. Les contextes sont déterminés à l'aide d'un algorithme de classification ou de *clustering*, puis leurs caractéristiques ainsi que l'ensemble des images de la collection sont utilisés pour aider à la détection des photographies de mauvaise qualité. Pour illustrer leurs travaux, l'exemple de la netteté est présenté dans les articles cités et il est démontré que le réseau élimine bien les photographies dont la netteté est jugée insuffisante dans le contexte de la collection. Plus récemment dans [Huang 20], une nouvelle architecture à base de réseaux convolutifs permet de regrouper des caractéristiques multi-échelles provenant de différentes couches du réseau.Effectivement, de manière intuitive, les caractéristiques générées par les couches convolutives supérieures sont riches en sémantique mais présentent une résolution faible, ce qui peut entraîner une perte de détails dans les images. C'est pour lever cette limite que l'approche multi-échelle est utilisée.

1.3. Résumé de collection d'images

Dans le domaine de la sélection de photographies au sein de séquences, les images sont souvent corrélées car elles sont prises dans un même lieu ou lors d'un même événement, ce qui signifie qu'elles présentent généralement des similitudes de contenu ou de contexte. L'objectif est de choisir la ou les meilleures images parmi cette série pour représenter de manière complète et significative l'événement ou le lieu. En revanche, dans le contexte du résumé de collection d'images ou *image summarization*, les images peuvent provenir de contextes différents et ne sont pas nécessairement corrélées, comme l'illustre la collection d'images présentée dans la Figure II.4. L'objectif est donc de condenser l'information visuelle à partir d'une collection d'images plus vaste en extrayant les éléments les plus importants ou les plus représentatifs parmi tous les événements ou lieux conte-

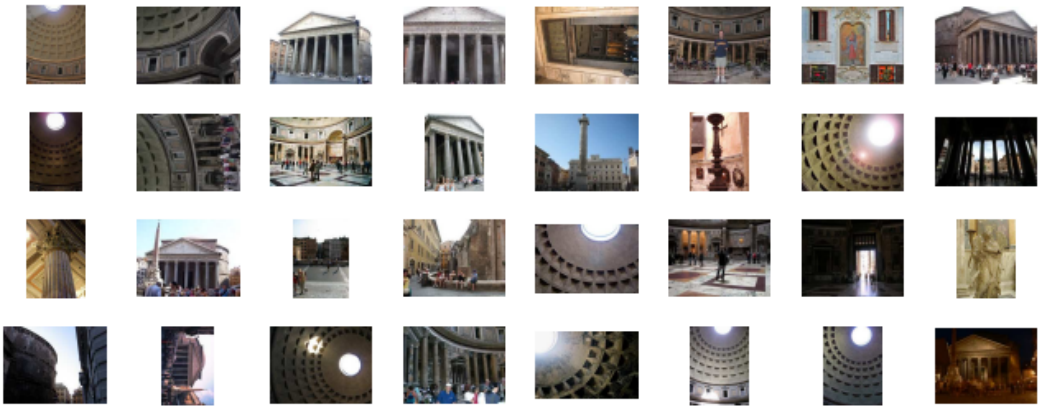


FIGURE II.4. – Exemple illustrant une collection aléatoire de 32 images non triées, extraites d'une vaste base de données comprenant des milliers d'images du Panthéon, comme présenté et illustré dans [Simon 07]. L'objectif principal est de sélectionner un sous-ensemble de vues canoniques pour servir de résumé. Bien que cet exemple ne montre qu'un seul monument, les bases de données, en général, contiennent des milliers d'images de plusieurs monuments, chacun devant être pris en compte dans le résumé.

nus dans la collection initiale. Ces techniques peuvent être appliquées à divers contextes, comme la sélection automatique des meilleures photographies d'une ville à partir d'une multitude d'images qui sont, soit géo-référencées et caractérisées par leur emplacement spatial [Jaffe 06], soit étiquetées à l'aide de mot clés [Kennedy 08]. D'après les travaux de [Sinha 11], un sous-ensemble d'images idéal à sélectionner doit respecter trois critères fondamentaux pour être le plus efficace possible : la qualité s'appuyant les attributs de l'image, la diversité et la couverture. La qualité des images sélectionnées dans le sous-ensemble reflète leur intérêt ou leur attrait global. La diversité, quant à elle, mesure le degré de non-redondance entre les images. Enfin, la propriété de couverture garantit que les concepts importants présents dans l'ensemble initial d'images sont également représentés de manière adéquate dans le sous-ensemble sélectionné. Nous avons distingué trois catégories d'approches que nous présentons dans la suite de cette section : les techniques s'appuyant sur des méthode de classification, celles s'appuyant sur un apprentissage profond, et enfin les approches qui sont spécifiques à l'usage de vidéos.

1.3.1. Résumé par classification

Certaines approches reposent sur des méthodes de classification pour partitionner des images similaires, afin de sélectionner ensuite la meilleure image dans chaque classe de la partition [Stan 03, Simon 07].

Dans l'étude menée dans [Sharma 15], chaque image est représentée par le vecteur obtenu après l'application du détecteur SIFT, *Scale Invariant Feature Transform*, proposé

par [Lowe 04] et célèbre en vision par ordinateur. Ce sont ces vecteurs qui sont exploités pour réaliser la classification. Pour créer ces partitions, certaines méthodes de classification utilisent des techniques basées graphes [Cai 04, Gao 05] pour extraire les caractéristiques présentes dans chaque image ou pour illustrer les liens sémantiques entre images [Riahi Samani 20]. Dans [Chen 00], les graphes sont utilisés sous forme d'arbre, et plus précisément sous forme de pyramide de similarité. Les auteurs proposent le concept de pyramide de similarité pour représenter les collections d'images. Chaque niveau est organisé de manière que les images similaires soient proches les unes des autres sur une grille 2D. Chaque image est associée à un vecteur de caractéristiques qui contient les informations pertinentes nécessaires pour mesurer la similarité entre les images, telles que les caractéristiques de couleur ou de texture. Les images sont d'abord organisées en un arbre binaire par le biais d'une classification exploitant les similarités par paire. L'arbre binaire est ensuite transformé en un arbre dans lequel chaque nœud a quatre enfants au lieu de deux, autrement dit sous la forme de *quadtree* ou quadrant. Un *quadtree* offre non seulement un plus grand choix à chaque niveau de la hiérarchie, mais il utilise également mieux l'espace d'affichage 2D. Les représentants des classes sont disposés de manière à maximiser la cohérence visuelle globale.

FIGURE II.5. – Exemple de graphes de scène utilisés dans [Pasini 22]. L'illustration est extraite de cette publication. Les objets sont étiquetés avec leur classe. La connaissance des relations entre ces différentes classes sémantiques permet de construire les graphes présentés dans ces trois images.

Les auteurs et les autrices de [Pasini 22] introduisent SImS, *Semantic Image Summarization*. Contrairement aux approches antérieures, leur méthode représente la collection d'images à l'aide de caractéristiques sémantiques représentant les classes d'objets et leurs relations, sous la forme de graphes de scène, cf. Figure II.5. Plus précisément, la méthode SImS exploite des techniques d'extraction de sous-graphes fréquents. Il est important de préciser que les graphes de scène sont des modèles utilisés en Infographie en 3D ou en vision par ordinateur en 2D pour représenter les relations entre les objets d'une même scène ou image.

1.3.2. Résumé à l'aide d'apprentissage profond

De nombreuses approches font appel à une étape d'apprentissage et comme pour le cas de la sélection de séries photographiques, les approches les plus récentes s'appuient sur les techniques d'apprentissage profond par réseaux de neurones [Deng 07a, Sharma 22].

FIGURE II.6. – Exemple de résumé obtenu par apprentissage profond. Nous présentons l'illustration de la projection 2D de la partition des images réalisée dans [Camargo 09]. Les images sont disposées en fonction de leur ressemblance, et celles encadrées en rouge sont sélectionnées pour représenter l'ensemble des images qui leur sont similaires.

Les auteurs de [Camargo 09] ont proposé une méthode de *kernel alignement* ou alignement de noyaux, pour réaliser la synthèse de collections d'images impliquant des connaissances du domaine. La technique du *kernel alignement* est une méthode d'apprentissage automatique qui exploite le calcul de la similarité entre deux distributions de données en utilisant des noyaux [1]. Un algorithme d'ascension de gradient est mis en place. Au final, l'image sélectionnée pour le résumé correspond au point central de chaque classe, cf. la Figure II.6 où les images sélectionnées sont encadrées en rouge. Nous pouvons également citer les travaux de [Sreelakshmi 21] qui permettent de comparer une approche classique basée SVM, *Support Vector Machine* et nommée *Image Summarization using Clustering*, ISUC, avec une seconde méthode, intitulée *Image Summarization using Auto encoder*, ISUA. Cette dernière consiste à entraîner un réseau de neurones utilisant une architecture d'auto-encodeur, afin d'assigner une valeur représentative à chaque image de la collection. D'après l'analyse des résumés de collections obtenus, le modèle ISUA sélectionne les images les plus appropriées et les plus diversifiées.

1. Pour rappel, un noyau, en apprentissage automatique, est une fonction mathématique qui calcule la similarité entre deux éléments d'un espace donné.

1.3.3. Résumé adapté au cas des vidéos

Enfin, il est important de souligner que ce besoin de sélection d'images peut également concerner les vidéos. On parle alors de *Video summarization* ou résumé de vidéo. L'objectif visé est alors double : d'une part, sélectionner des images de qualité visuelle élevée mais également les plus représentatives, c'est-à-dire qui résume les informations les plus importants du contenu vidéo. Bien évidemment, comme il s'agit de vidéos, les temps de calculs peuvent être beaucoup plus longs et cet aspect est également considéré. Ainsi, le réseau convolutif introduit par [Ren 20] est conçu pour classer les images extraites de courtes séquences vidéo, afin de sélectionner automatiquement les meilleures. Chaque courte vidéo compte en moyenne 19 images et chaque image a été annotée manuellement, par des utilisateurs et utilisatrices en précisant un score indiquant à quel point une image est représentative de la vidéo. Avec ces annotations, le réseau de neurones, construit avec une architecture siamoise, est capable d'apprendre à classer, par comparaisons deux à deux, les images au sein d'une même vidéo. Enfin, dans [Singh 20], les auteurs ajoutent que contrairement à la vidéo, les images ne bénéficient pas de liens temporels exploitables par des réseaux neuronaux tels que les LSTM, *Long-Short-Term-Memory*, ou les RNN, *Recurrent Neural Network*. Ces réseaux sont plus complexes par nature mais permettent d'exploiter la redondance temporelle de ces données contrairement à leur approche antérieure basée réseaux de neurones adverses génératifs, GAN, *Generative Adversarial Networks* [Singh 19].

Enfin, d'après les résultats de [Zhu 20], les méthodes MSMO existantes sont principalement entraînées sur les modalités textuelles, ce qui crée un biais car la qualité de l'image sélectionnée par le modèle pendant l'entraînement n'est pas prise en compte. Ainsi, ils proposent une nouvelle fonction objective multimodale qui prend en considération à la fois la génération du résumé et la sélection de l'image.

1.4. Évaluation de la qualité esthétique d'image

Une approche alternative pour déterminer les images les plus pertinentes à conserver consiste à évaluer leur qualité esthétique. L'évaluation esthétique des images vise à distinguer de manière numérique les photographies de haute qualité de celles de faible qualité sur la base de règles photographiques, souvent à l'aide d'une classification binaire ou d'une estimation de score de qualité pour chaque image. Diverses approches ont été proposées dans la littérature pour tenter de résoudre ce problème à la fois subjectif et difficile et nous présentons ici un état de l'art qui reprend, résume et complète celui présenté dans les travaux de [Deng 17, Zhai 20, Zhang 21].

Les approches visant à évaluer la qualité esthétique procèdent à l'extraction de caractéristiques spécifiques de chaque image. Ces caractéristiques peuvent être de bas niveau

(utilisation d'un critère d'évaluation s'appuyant sur la couleur, la texture ou l'éclairage), telles que celles présentées dans la section 1.4.1 ou de haut niveau (composition, mémorabilité, contenu et esthétisme), cf. section 1.4.2.

1.4.1. Utilisation de caractéristiques bas niveau

Éclairage	Texture	Couleur
[Luo 08] [Wong 09] [Bychkovsky 11] [Kaufman 12] [Yuan 12] [Saha 15] [Fang 17] [Wang 19] [Guo 20]	[Datta 06] [Li 09] [Wong 09] [Tang 11] [Lo 12] [Saha 15] [Zhang 15] [Fang 17] [Zhang 16]	[Datta 06] [Li 09] [Nishiyama 11] [Lo 12] [Sartori 15] [Amirshahi 15] [Saha 15] [Kong 16] [Kim 19] [Zhan 19] [Jin 19] [Sheng 20] [Leder 22]

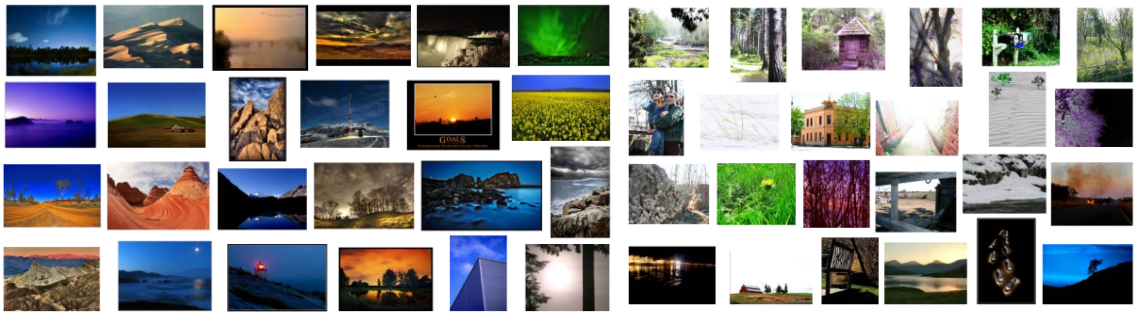
TABLE II.1. – Publications sur l'estimation de la qualité esthétique d'images en utilisant des caractéristiques bas niveau : évaluation de la qualité de l'éclairage, de la texture ou de la couleur.

Dans la Table résumée II.1, l'**éclairage** fait référence aux méthodes qui étudient la manière dont l'éclairage d'une scène, dans une photographie, peut modifier son ambiance et la perception des êtres humains qui la regardent. Dans les photographies considérée comme étant de haute qualité, l'éclairage fait en sorte que les sujets n'apparaissent pas plats et renforce leur impression de 3D afin que ces sujets attirent le regard. De plus, on juge qu'un éclairage est adapté ou utile s'il crée un fort contraste entre le sujet et l'arrière-plan. Ces approches s'appuient par exemple sur l'estimation de la luminosité moyenne dans une image [Yuan 12].

Les approches s'appuyant sur la **texture**, toujours listées dans la Table II.1 concernent les méthodes qui analysent le caractère granuleux ou lisse d'une photographie, pouvant indiquer la présence, ou l'absence de texture ainsi que sa nature. L'utilisation de la texture est une compétence liée au domaine de la composition, en photographie. L'une des manière de mesurer le caractère granuleux d'une image consiste à utiliser la transformée en ondelettes de Daubechies [Daubechies 92], qui a souvent été utilisée dans la littérature pour caractériser la texture, comme dans [Wong 09]. En général, l'analyse de la texture d'une image fournit des informations complémentaires à celle issue de l'analyse de la couleur et sont souvent combinées, comme dans [Lo 12].

En photographie, on estime qu'une grande partie de ce que les êtres humains perçoivent et ressentent d'une photographie passe par les couleurs. Les approches exploitant la **couleur** sont également listées dans la Table II.1. Certaine couleur ou une certaine combinaison de couleurs peuvent avoir une influence sur les émotions ou les sentiments humains, produisant ainsi une réponse affective agréable. Pour estimer la qualité d'une

image par rapport à ses couleurs, certaines méthodes étudient la combinaison des couleurs, c'est-à-dire la manière dont les couleurs sont organisées dans une image [Sartori 15], tandis que d'autres approches se concentrent sur l'harmonie des couleurs, en examinant comment les différentes couleurs se combinent pour créer une vision esthétique agréable [Nishiyama 11, Kong 16], cf. Figure II.7.



(a) Qualité esthétique élevée.

(b) Qualité esthétique faible.

FIGURE II.7. – Exemples de photographies classées selon leur qualité colorimétrique [Nishiyama 11]. L'illustration extraite de la publication présente d'une part, les images de bonne qualité, en (a), et d'autre part, celles de mauvaise qualité, en (b).

1.4.2. Utilisation de caractéristiques haut niveau

Composition	Contenu	Mémorabilité	Esthétisme
[Luo 08] [Li 09] [Bhattacharya 10] [Liu 10] [Dhar 11] [Guo 12] [Park 12] [Zhang 12] [Amirshahi 15] [Jin 19] [Leder 22]	[Dhar 11] [Kaufman 12] [Tang 13] [Amirshahi 15] [Kong 16] [Inoue 18] [Lu 20] [Achlioptas 21]	[Leder 04] [Dubey 15] [Isola 11] [Khosla 14] [Peng 16] [Cetinic 19] [Serre 19] [Cornia 20]	[Li 09] [Dhar 11] [Murray 12] [Lu 14] [Amirshahi 15] [Lu 15] [Kong 16] [Kao 17] [Ren 17] [Schwarz 18] [Jin 19] [Viswanatha Reddy 20] [Xu 20] [He 22]

TABLE II.2. – Publications sur l'estimation de la qualité esthétique d'images en utilisant des caractéristiques haut niveau en faisant intervenir les concepts de composition, d'intérêt du contenu, de mémorabilité et d'esthétisme.

Dans la Table résumée II.2, la notion de **composition** fait référence à l'organisation de tous les éléments visuels à l'intérieur d'une photographie. Une bonne composition doit permettre de montrer clairement le sujet de la photographie. On cherche à quantifier ou à évaluer le contraste entre la lumière et l'obscurité, entre les formes et les couleurs. Deux

des principes de composition photographique les plus connus sont : la règle des tiers [2] et le nombre d'or [3] . Certaines approches étudient la composition des images à l'aide de cartes de saillance qui sont fortement corrélées aux directives de composition de la photographie puisqu'elles représentent les emplacements des objets saillants [Park 12].

Dans la Table II.2, nous parlons d'utilisation du **contenu** lorsque la sélection s'effectue sur la base d'analyse du respect des règles de photographie, en fonction do contenu photographié cf. Figure II.8. Par exemple, pour les photographies en gros plan, les photographes privilégient un contraste élevé entre le premier plan et l'arrière-plan. Pour les portraits humains, une attention particulière est portée sur les réglages de l'éclairage pour créer des motifs esthétiquement agréables sur les visages humains. Pour les photographies de paysage, une structure spatiale bien équilibrée est favorisée. Pour exploiter les informations relatives au contenu de l'image, telles que les catégories de scènes ou le choix du sujet photographique, dans [Tang 13], les auteurs proposent de segmenter les régions en fonction de la taille de l'image et d'extraire les caractéristiques visuelles sur la base de la catégorisation du contenu des photographies.

(a) Catégorie Nuit.

(b) Catégorie Architecture.

(c) Catégorie Paysage.

(d) Catégorie Humain.

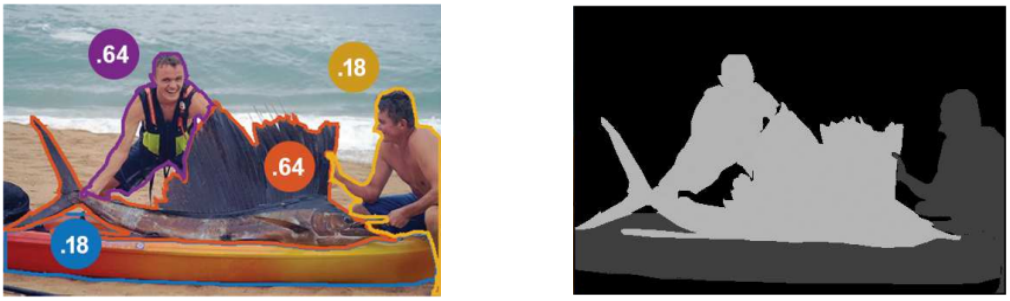
FIGURE II.8. – Estimation de la qualité des images en fonction de leur contenu. Nous illustrons les résultats proposés dans la publication de [Tang 13]. La première ligne présente les photographies de haute qualité alors que la seconde ligne expose les photographies de faible qualité.

Les méthodes analysant pourquoi certaines images restent gravées facilement dans la mémoire, alors que d'autres sont ignorées ou rapidement oubliées concernent la notion de **mémorabilité**. Ainsi, les publications listées dans la Table II.2 analysent divers facteurs visuels susceptibles d'influencer la mémorisation d'un objet (par exemple, la couleur, la saillance visuelle et les catégories d'objets), mais également la corrélation entre la mémo-

2. La règle des tiers s'appuie sur la division de l'image en neuf parties égales à l'aide de deux lignes horizontales et deux lignes verticales, espacées de manière uniforme. Cela crée quatre , créant ainsi des points stratégiques au niveau des intersections, où on suppose que l'œil est attiré naturellement.

3. Le nombre d'or en photographie est un principe esthétique utilisant le rapport de proportion de 1 : 1.618, considéré comme visuellement agréable. On suppose que les éléments clés de la scène doivent être placés en respectant ce rapport.

rabilité de l'objet et celle de l'image. Généralement, la mémorabilité est liée à un score d'importance attribué à chaque région segmentée de la scène, cf. Figure II.9. La mémorabilité d'une image est fortement influencée par la mémorabilité de son objet le plus mémorable. Certains travaux, comme [Isola 11], ont remarqué que les images contenant des personnes ont tendance à être plus mémorables que d'autres.



(a) (b)

FIGURE II.9. – Illustrations, en (a) des scores de mémorabilité de chaque objet présent dans l'image, obtenus lors de l'étude utilisateur et utilisatrice réalisée dans [Dubey 15]. Certains objets, comme le poisson et la personne de gauche, sont plus mémorables que d'autres. L'image segmentée, en (b), montre la carte de la vérité terrain générée à partir des segments d'objets et des scores de mémorabilité.

L'essor de l'apprentissage, et en particulier de l'apprentissage profond, a largement contribué à l'émergence des approches s'appuyant sur la notion d'**esthétisme**, dernière catégorie de publications listées dans la Table II.2. Lors de l'entraînement, ces scores peuvent provenir d'études utilisateur et utilisatrice, impliquant des photographes ou des individus non spécialistes, ainsi que des bases de données telles que *AVA* dans [Murray 12]. Les modèles sont généralement formés soit pour effectuer une classification binaire ou une régression pour prédire un score, soit pour produire des classements d'images, selon qu'ils visent à estimer la qualité esthétique globale d'une image ou à sélectionner l'image la plus esthétiquement plaisante parmi plusieurs. Souvent, la qualité esthétique d'une image correspond à la combinaison de plusieurs critères, tels que ceux mentionnés dans les catégories précédentes. Par exemple, dans [Ren 17], la qualité esthétique d'une image correspond à la combinaison de critères tels que la règle des tiers et l'harmonie des couleurs.

1.5. Résumé et analyse des approches d'estimation de la pertinence en 2D

Toutes les approches mentionnées abordent une problématique proche de notre problématique. Toutefois, nous n'envisageons pas une sélection manuelle ou semi-manuelle comme dans les travaux de [Jacobs 10]. De plus, certaines de ces approches traitent des

images avec des redondances significatives [Chang 16, Ren 20], tandis que nous ne possédons que des images faiblement corrélées représentant un même objet dans divers environnements, avec une faible quantité d'information commune. De plus, toutes ces approches exploitent uniquement des images, ce qui limite la qualité estimée à une évaluation globale de l'image, sans considération pour un objet cible spécifique. En effet, dans ce travail de thèse, nous supposons que nous possédons un modèle 3D dont nous pouvons extraire des informations supplémentaires, telles que la géométrie de l'objet, indépendamment de la texture, de la couleur et de l'éclairage.

En résumé, par opposition aux approches que nous venons de décrire, nous souhaitons évaluer la qualité de l'image par rapport à un objet spécifique qu'elle représente, en tenant compte de sa position, de son orientation et de sa visibilité. Autrement dit, nous souhaitons évaluer l'intérêt d'une vue 2D d'un objet par rapport à son modèle 3D, en tirant parti des informations géométriques que nous pouvons extraire de cette vue. Cela signifie que nous souhaitons proposer un classement des vues 2D en fonction de la qualité de la mise en valeur de l'objet par l'image, plutôt qu'en fonction de la qualité intrinsèque de l'image elle-même. Pour réaliser cette évaluation de la pertinence, nous jugeons nécessaire d'extraire les informations les plus pertinentes offertes par le point de vue de l'image et d'examiner si elles correspondent à celles de l'objet 3D. Nous avons donc besoin d'outils pour extraire cette information en 2D et en 3D et la section suivante va nous permettre de présenter en détail les approches de l'état de l'art qui traitent de la détection de points caractéristiques en 2D et en 3D.

2. Détection de points saillants en 2D et 3D

Nous désignons par le terme point d'intérêt tout point pouvant être détecté et extrait d'une image car il possède des caractéristiques saillantes qui le distinguent des autres points de la scène. Ces points se caractérisent par des attributs particuliers tels qu'un fort contraste, une texture marquée ou des couleurs nettement distinguables du reste de la scène. Les outils proposés pour détecter ces points d'intérêt sont des détecteurs et nous proposons de reprendre la classification proposée dans [Chambon 20] pour les présenter, à savoir : les approches basées premier ordre, basées régions, basées second ordre et enfin les approches multi-échelle.

Comme en 2D, un point d'intérêt 3D va correspondre à un élément saillant, un élément caractéristique en 3D. La définition d'un point d'intérêt en 3D dépend de la nature des données 3D manipulées. En effet, cela peut être une carte de profondeur (associé à un point de vue), un nuage de point 3D (sans notion de topologie), un maillage ou encore un modèle continu, explicite ou implicite. Nous parlons de données structurées (volumes/voxels, maillage) ou de données non structurées (nuages de points). L'information disponible peut donc être fondamentalement [Bustos 09]. Cependant, nous pouvons retrouver les mêmes

familles de détecteurs qu'en 2D. En effet, la plupart des techniques s'appuient sur la généralisation des détecteurs du 2D au 3D. Cependant, en 3D les approches du second ordre sont nettement favorisées par rapport aux autres à cause de leur utilisation de la courbure qui intuitivement est plus adaptée à l'étude des surfaces en 3D. Ainsi, nous allons présenter les cinq familles de détecteurs que nous avons énumérées en présentant les outils disponibles en 2D et en 3D.

Dans la suite de ce document, nous aborderons régulièrement la notion de **répétabilité**. Il s'agit de la capacité d'un détecteur de point d'intérêt à détecter le même point physique de la scène, quelle que soit l'image ou la modalité (2D/3D) dans laquelle il est visible.

2.1. Approches basées premier ordre

En 2D, il existe de nombreux détecteurs et il est classique d'utiliser des détecteurs du premier ordre qui s'appuient sur les dérivées premières des images, autrement dit sur le calcul des gradients. Dans cette catégorie, nous retrouvons un des détecteurs les plus connus et les plus anciens, celui de [Harris 88]. Le principe est d'estimer à quel point un point étudié est corrélé à ses voisins, en s'appuyant sur la publication initiale de Moravec et al. [Moravec 80]. Ils analysent le comportement des valeurs propres de la matrice des dérivées premières. Mais ces techniques ne tiennent pas compte des difficultés dues aux textures, aux changements de lumière ou aux changements d'échelle. Pour obtenir un détecteur de points d'intérêt 3D, dans [Sipiran 11], le détecteur de Harris a été adapté en 3D. Il existe également des travaux comme ceux de [Hafiz 15] utilisant des opérateurs de Sobel-Harris, ou encore l'approche décrite dans [Ahmed 18], faisant intervenir un algorithme s'appuyant sur celui du *mean-shift*, pour détecter des coins et des frontières.

2.2. Approches basées région

Pour ce type de détecteur, le principe commun est de considérer non plus un point caractéristique comme un point qui se distingue des autres, mais plutôt, comme le centre d'une région qui se distingue des autres. Au final, la réponse renvoyée est un point, mais, la notion rapportée est plutôt celle d'une région. Nous pouvons citer les détecteurs connus comme [Smith 97, Matas 02, Tuytelaars 04, Rosten 06]. Dans [Tuytelaars 06], l'objectif est de trouver une nouvelle méthode de représentation de l'image permettant de décrire ses objets sans recourir à une segmentation. En conséquence, diverses versions de détecteurs locaux, utilisant l'analyse de régions d'intérêt autour d'un point, appelées *blob*, ont été introduites. Dans [Matas 02], les composantes connexes d'une image seuillée sont utilisées par leur détecteur nommé MSER, *Maximally Stable Extremal Regions*. Le détecteur IBR, *Intensity Based Regions* [Tuytelaars 04], détermine des régions d'intérêt autour des points en prenant en compte les maxima le long de droites partant du point étudié. La région retenue est obtenue en calculant une approximation par une ellipse de la région formée

par tous ces maxima locaux. Plusieurs détecteurs utilisant des comparaisons d'intensité ont été proposés en comparant l'intensité des pixels des régions environnants avec celle du pixel central afin de simplifier le calcul du gradient de l'image, effectué dans les détecteurs du premier ordre. L'approche de SUSAN, *Smallest Univalue Segment Assimilating Nucleus* [Smith 97] examine la proportion de pixels photométriquement proches du pixel étudié dans une région circulaire autour de celui-ci. Une approche similaire, mais plus simple à mettre en œuvre et offrant une accélération des calculs, est FAST, *Features from Accelerated Segment Test* [Rosten 06]. Dans une version améliorée du détecteur FAST, introduite dans [Rosten 08] et nommée FAST-ER, *Enhanced Repeatability*, la fonction de coût a été ajustée pour inclure un terme lié à la répétabilité des points d'intérêt détectés. L'objectif est de maximiser le nombre de points d'intérêt répétables dans une paire d'images. Une autre amélioration est AGAST, *Adaptive and Generic Accelerated Segment Test* [Mair 10], dans laquelle deux critères supplémentaires de comparaison de la luminosité des pixels sont définis, rendant ainsi le détecteur FAST plus adaptatif. Aucune approche 3D ne réalise de recherche de régions d'intérêt comme dans ces travaux 2D.

2.3. Approches basées second ordre

Enfin, une dernière famille de détecteurs sont ceux utilisant les dérivées secondes et, plus précisément, la notion de courbure. Ces détecteurs du second ordre exploitent donc les informations provenant de la matrice hessienne en raison de son indépendance vis-à-vis des changements d'ordre zéro et d'ordre un, ainsi que de ses bonnes performances en termes de temps de calcul et de précision. Une des plus anciennes approches est celle utilisant l'opérateur Hessien [Beaudet 78], elle estime la courbure de la surface alors que le détecteur de Kitchen et Rosenfeld [Kitchen 82] s'appuie, pour un point donné, sur la notion de courbure du contour passant par le point considéré. Par ailleurs, à chaque point utilisé est assigné un poids en fonction de la norme du gradient, car les auteurs veulent tenir compte des situations dans lesquelles la norme du gradient est faible (indiquant ainsi un contour peu contrasté, pouvant correspondre à un bruit). Le principe de [Fischer 14] est de déterminer le changement de gradient image le long de la tangente pour obtenir un scalaire q approchant κ, la courbure le long de la surface. Dans le même principe, le détecteur PCBR [Deng 07b], *Principal Curvature-Based Regions*, propose une technique qui exploite soit la valeur propre minimale, soit la valeur propre maximale de la matrice hessienne dans l'objectif de trouver la courbure principale. En effet, l'utilisation de la valeur propre minimale permet de mettre en valeur une frontière sombre sur un fond clair alors que la maximale met en valeur une frontière claire sur un fond sombre. Enfin, le détecteur nommé Saddle [Aldana-Iuit 16] extrait les points dont les voisinages, lorsqu'ils sont traités comme une surface d'intensité 3D, présentent des profils concaves et convexes dans une paire de directions orthogonales, cf. Figure II.10. Le détecteur Saddle peut atteindre une

plus grande répétabilité et une plus grande répartition que les méthodes traditionnelles, et que certaines méthodes d'apprentissage. Ces observations, ont été réalisées dans les travaux de [Komorowski 18] en réalisant une étude comparative entre les détecteurs de points d'intérêt récents, basés apprentissage profond.

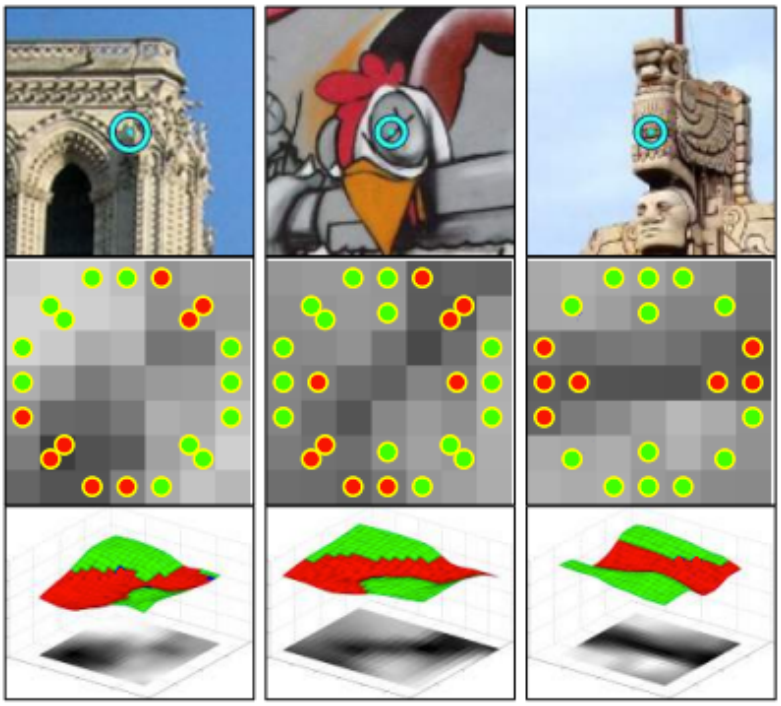


FIGURE II.10. – Illustration du détecteur Saddle basé second ordre dans [Aldana-Iuit 16]. En haut, il s'agit d'exemples de points d'intérêt, visualisation des voisinages dans l'image avec en rouge (respectivement vert) les zones d'intensité convexes (concaves), et les représentations 3D des intensités.

En 3D, la méthode ISS, *Intrinsic Shape Signatures* [Zhong 09], utilise les variations locales de la courbure, le long de chaque direction principale de la surface dans les nuages de points pour détecter uniquement les points d'intérêt présentant des changements de courbure significatifs. Plus précisément, la saillance de chaque sommet est dérivée de la décomposition des valeurs propres de sa matrice de covariance non-normalisée. Ensuite, le rapport des valeurs propres est utilisé pour sélectionner certains points, et la saillance finale est déterminée par le vecteur propre. De manière classique et proche de ce qui se fait en 2D, les auteurs de [Shaiek 12] ont proposé une méthode qui utilise les deux courbures principales, en combinant la courbure moyenne et la courbure gaussienne.

2.4. Notion d'analyse multi-échelle

Suivant l'échelle d'analyse choisie, nous détectons des détails caractéristiques différents. En effet, avec une échelle faible, seuls des détails fins sont considérés et ainsi des détails non

FIGURE II.11. – Illustrations de l'introduction d'une analyse multi-échelle dans une pyramide d'images (multi-résolution), extraites de [Gales 11]. De haut en bas, la résolution des images est diminuée, formant ainsi une pyramide d'images. De gauche à droite un filtrage gaussien est appliqué avec un filtre de plus en plus grand, pour fournir une analyse multi-échelle.

significatifs peuvent être pris en compte et produire des détections non pertinentes alors qu'avec une échelle trop importante, des détails fins mais pertinents de l'image peuvent être ignorés. Ainsi, pour palier cette difficulté du choix de l'échelle, les généralisations du détecteur d'Harris pour donner l'approche Harris-Laplace et de l'opérateur Hessien, pour créer l'approche multi-échelle Hessien-Laplace ont été introduites dans [Mikolajczyk 04]. Dans de nombreux cas, cette capacité à s'adapter aux différentes échelles va se traduire par le fait de faire varier un (ou plusieurs) paramètre(s) utilisés par le détecteur. Par exemple, dans le cas du célèbre détecteur SIFT, *Scale Invariant Features Transform* [Lowe 04], c'est la taille du filtre gaussien utilisé pour le lissage et pour le calcul du laplacien qui va varier. Cette variation permet de mettre en évidence les détails plus ou moins fins des données étudiées. Plus précisément, la méthode SIFT, s'appuie sur la détection des extrema en échelle et en espace du laplacien. Les auteurs utilisent des différences de gaussiennes pour approximer le laplacien et ils effectuent les calculs avec différentes tailles (échelles) de filtre dans différentes résolutions (utilisation d'une pyramide d'images). Selon les conclusions de [Morel 09a], les filtres gaussiens sont les filtres linéaires permettant un lissage et une analyse multi-échelle la plus efficace. Le détecteur SURF [Bay 08], *Speeded Up Robust Features*, s'appuie également sur l'analyse de la matrice hessienne, et est présenté comme étant plus rapide que SIFT et d'autres techniques multi-échelles en utilisant une approximation du Laplacien et des astuces algorithmiques. Grâce à leurs grandes performances, ces deux détecteurs multi-échelle ont, eux aussi, de nombreuses extensions en 3D. Pour le détecteur SIFT, nous pouvons trouver 3D SIFT [Scovanner 07, Flitton 10, Sattler 11] et PointSIFT [Jiang 18], une version 3D basée apprentissage. En ce qui concerne le détecteur SURF, il existe 3D-SURF [Knopp 10]. De nombreuses améliorations de SIFT ou SURF, telles que l'amélioration de l'invariance par rapport aux transformations af-

fines [Morel 09b] et de la répétabilité [Mainali 13], ont été proposées. En 2012, les auteurs de [Alcantarilla 12] ont présenté un nouveau détecteur nommé KAZE [4] qui exploitent les filtres de diffusion non linéaire, contrairement aux filtres gaussiens largement utilisés dans les pyramides d'images. Ce détecteur s'inspire largement des mêmes principes que SIFT et SURF avec un filtrage multi-échelle. Les auteurs ont également proposé une version accélérée dans [Alcantarilla 13], nommée Accelerated-KAZE. Selon leurs observations, les détecteurs SIFT et SURF s'avèrent être les plus invariants à l'échelle (sur la base de la répétabilité). De plus, les détecteurs SIFT, KAZE et AKAZE ont une plus grande précision lorsque des rotations ont été appliquées aux images. Finalement, la précision globale de SIFT est la plus élevée pour tous les types de transformations géométriques et SIFT est alors désigné comme l'algorithme le plus précis.

De même, en ce qui concerne les détecteurs basés régions, les auteurs de [Tuytelaars 04] ont proposé une version multi-échelle avec l'opérateur EBR, *Edge-Based Regions*, qui étudie les familles de parallélogrammes définis à partir du point étudié et des points de contours passant par le point étudié. Dans la famille des détecteurs basés courbure, il existe le détecteur multi-échelle CSS [Mokhtarian 03], *Curvature Scale Space*, qui recherche les extrema de la courbure le long des contours en utilisant une approche multi-échelle.

Enfin, certains travaux réalisant un extraction de points d'intérêt en 3D, peuvent aussi proposer des techniques multi-échelle comme les travaux de [Nader 14] ou [Lei 17], qui s'appuient sur l'extraction de caractéristiques de la surface, comme la courbure.

2.5. Approches basées apprentissage

Historiquement, il existe des détecteurs de points d'intérêt qui font intervenir une étape d'apprentissage. Le modèle TaSK [Strecha 09], *Task Specific Keypoint* vise à améliorer la répétabilité des points d'intérêt en apprenant à post-filtrer les points d'intérêt ayant les caractéristiques les plus stables. Il s'agit des points qui peuvent être appariés de manière fiable sur de multiples images et qui possèdent des descripteurs suffisamment distinctifs par rapport à ceux des pixels voisins. Dans leurs travaux, le détecteur utilisé est celui de [Förstner 87]. Le modèle TILDE [Verdie 15], *Temporarily Invariant Learned Detector*, vise à détecter des points d'intérêt répétables dans le cas où nous observons des changements importants de la scène, comme par exemple, des changements météorologiques ou des variations d'éclairage. L'ensemble d'entraînement se compose de plusieurs piles d'images prises du même point de vue, mais à des saisons et heures différentes dans lesquelles le détecteur SIFT [Lowe 04] a été appliqué. Le modèle est entraîné pour calculer une carte de scores, c'est-à-dire une valeur pour chaque région, de taille fixe, de l'image d'entrée. Ainsi, la fonction de perte est construite pour que le modèle fournisse une valeur élevée pour les régions dites positives (régions centrées près de l'emplacement d'un bon

4. En référence au vent qui se nomme ainsi en japonais.

point d'intérêt) et une petite valeur sur les régions dites négatives.

La première méthode s'appuyant sur l'apprentissage profond, avec une architecture de type siamois est LIFT, *Learned Invariant Feature Transform* [Yi 16]. Elle permet la détection de caractéristiques, l'estimation d'orientation et l'extraction robuste de descripteurs. Son entraînement s'appuie sur des points d'intérêt fournis par la méthode SIFT [Lowe 04] et il est validé par une approche de type SfM, *Structure from Motion*. Des réseaux de neurones convolutifs ont également été utilisés pour améliorer la détection de points d'intérêt dans les images. Par exemple, le réseau proposé par [Altwaijry 16] apprend simultanément à détecter des points d'intérêt et à calculer les descripteurs des régions situées autour de ces derniers, cf. Figure II.12. Sa particularité repose sur sa fonction de perte qui est composée de deux termes : un terme de classification binaire pour choisir des régions centrées autour de l'emplacement du point d'intérêt et un deuxième terme pour pénaliser les réponses du réseau sur les régions non centrées. Comme le modèle LIFT, l'ensemble d'entraînement est construit à l'aide d'un détecteur de points d'intérêt classique et validé par une approche de SfM, *Structure from Motion*.

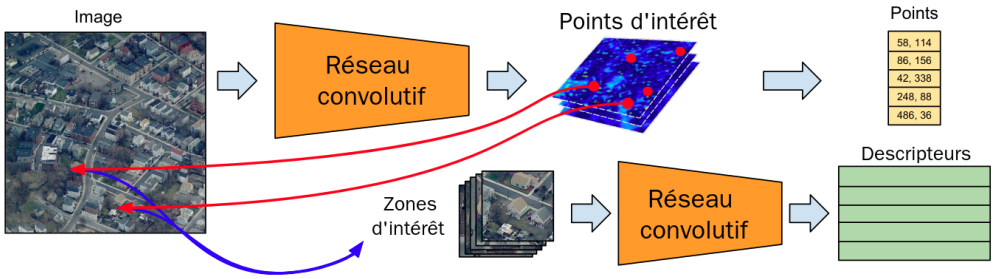


FIGURE II.12. – Réseau convolutif pour la détection de points d'intérêt dans des images. Illustration extraite de [Altwaijry 16]. Cette architecture proposée produit une pyramide d'échelles de réponses indiquant la présence de points d'intérêt. Ces cartes de réponses sont utilisées pour extraire des zones d'intérêt, auxquelles le réseau associe un descripteur.

Les Quad-networks [Savinov 17] correspondent à une catégorie de méthodes non supervisée de détection de points d'intérêt, qui ne reposent pas sur un détecteur de caractéristiques classique pour la génération de l'ensemble d'apprentissage. Le réseau est entraîné à cartographier l'image d'entrée en attribuant un score à chaque point, puis à classer les points en fonction de la réponse sous forme de carte de chaleur. Superpoint [DeTone 18] est un réseau auto-supervisé entraîné pour la détection des points d'intérêt et l'apprentissage des descripteurs. Il calcule conjointement les emplacements des points d'intérêt et les descripteurs associés. Plus précisément, un réseau convolutif de type VGG-*style* [Simonyan 14] est utilisé pour réduire la taille de l'image d'entrée. Le réseau se divise ensuite en deux parties : un décodeur de points d'intérêt et un décodeur de descripteurs. Le détecteur de points d'intérêt est pré-entraîné sur des données synthétiques constituées d'un grand ensemble d'images générées par ordinateur avec des emplacements de points d'in-

térêt correspondant à une pseudo-vérité du terrain. Ensuite, le réseau LF-Net [Ono 18], *Local Feature Network*, opte pour une architecture différente : une seule branche est considérée pour optimiser l'entrainement. Il utilise également un réseau de neurones convolutif sur des images pour générer une carte de chaleur, autrement dit, il attribue à chaque pixel un score en fonction de ses caractéristiques. Cette carte est ensuite utilisée pour extraire les points d'intérêt et leurs attributs, tels que l'échelle et l'orientation. Elle permet également l'augmentation de la précision et l'amélioration de la saillance des points d'intérêt. Similaire à LF-Net, RF-Net [Shen 19], *Receptive Field Network*, sélectionne des pixels possédant une forte réponse sur de multiples échelles, comme points d'intérêt, pour produire des cartes de chaleur. Les auteurs de [Bhowmik 20] proposent une approche différente pour la sélection des points d'intérêt : utiliser les principes de l'apprentissage par renforcement pour résoudre cette tâche de nature discrète. Enfin, ASFeat [Luo 20] est une approche visant à analyser les informations sur la forme locale des points caractéristiques afin d'améliorer la précision de la détection. Cette tâche a été réalisée en effectuant un apprentissage conjoint des détecteurs de caractéristiques et des descripteurs locaux.

Certaines méthodes combinent des caractéristiques issues de méthodes classiques avec celles fournies par des méthodes d'apprentissage pour la détection des points d'intérêt. Par exemple, le réseau KeyReg [Kim 21] utilise une succession de régresseurs s'appuyant sur le principe des forêts aléatoires pour déterminer des candidats répétables et fiables. Pour augmenter la répétabilité des points d'intérêt, les modules de forêt aléatoire sont appliqués à des images multi-échelles, et les points d'intérêt prédits à chaque échelle reçoivent un score de confiance. Chaque point candidat est détecté comme point d'intérêt final par un processus de suppression des non-maxima en s'appuyant ce score de confiance. De même, les auteurs de [Barroso-Laguna 22] utilisent, dans leur réseau de neurones convolutif Key.Net, des caractéristiques classiques comme structures d'ancrage fournies aux différentes couches convolutives du réseau, qui localisent, notent et classent les caractéristiques répétables. La représentation multi-échelle, ainsi que la fonction de perte utilisée, permettent de détecter les points robustes qui sont localisés sur plusieurs échelles.

Pour détecter et extraire des points d'intérêt sur des données 3D, les auteurs de [Streiff 21] utilisent un réseau d'extraction de caractéristiques 2D comme base de leur réseau appelé 3D3L, qui exploite à la fois l'intensité et la profondeur des images LiDAR pour extraire de précieuses caractéristiques 3D. Le réseau de neurones *Unsupervised Stable Interest Point* [Li 19] est un détecteur de points d'intérêt 3D utilisant de l'apprentissage non supervisé. Plus précisément, il peut détecter des points d'intérêt reproductibles et les localiser avec précision à partir de nuages de points 3D soumis à des transformations arbitraires. Plus récemment, le réseau décrit dans [Li 23] est capable de détecter des points d'intérêt 3D et d'estimer leurs descripteurs. Plus précisément, les points sont classifiés en fonction de leur courbure et un échantillonnage adaptatif est appliqué pour obtenir des points d'intérêt candidats. En outre, leur méthode prend en considération les positions géométriques des

points d'intérêt candidats et les variations des caractéristiques géométriques des points voisins.

2.6. Points répétables en 2D et 3D

Dans cette thèse nous avons à notre disposition un objet 3D, fourni avec son maillage 3D, et un ensemble d'images de cet objet. Pour extraire et comparer l'information essentielle de l'objet dans les diverses modalités utilisées, il est nécessaire d'avoir un détecteur permettant de récupérer des éléments saillants répétables, aussi bien en 2D qu'en 3D. Cette problématique est étudiée dans le domaine de l'analyse et de la compréhension d'images médicales, ainsi que dans la robotique et la reconnaissance biométrique, où il est souvent nécessaire de mettre en correspondance des données de natures différentes. Cette mise en correspondance, qui comprend une étape de détection de points d'intérêt, est nécessaire pour des applications telles que la détection précise des zones à traiter dans le domaine médical pour établir le meilleur diagnostic. La suite de cette section correspond à l'état de l'art présenté dans [Chambon 20].

Pour la robotique, la correspondance 2D/3D est très importante pour de nombreuses tâches qui nécessitent de déterminer la position 3D d'un objet d'intérêt : la publication [Pomerleau 15] propose une synthèse de l'état de l'art actuel de ce domaine. Dans le domaine de la reconnaissance biométrique, nous pouvons citer des travaux en reconnaissance de visages de [Yang 08]. L'appariement direct entre une image 2D et un modèle 3D peut être également utilisé pour localiser précisément le lieu d'une photographie [Irschara 09].

Dans la littérature, il existe également des travaux traitant de la même problématique de mise en correspondance, mais dont les données d'entrée sont différentes de celles que nous utilisons. Par exemple, dans [Toshev 09], des objets sont reconnus dans des vidéos, à partir de modèles 3D connus. Une segmentation du mouvement permet d'extraire chaque objet. Ensuite, la silhouette de l'objet est utilisée pour réaliser l'appariement. L'avantage de cette approche est que les auteurs possèdent de nombreux points de vue sur l'objet à reconnaître. Dans notre travail, nous nous intéressons spécifiquement à la reconnaissance d'objets à partir d'une image 2D texturée et d'un modèle 3D non texturé. Pour comparer ces données, il est possible de conserver les différentes modalités et de faire une mise en correspondance directe entre les deux modalités de nature différentes. De manière très classique, pour effectuer cette tâche, des points d'intérêt peuvent être détectés, caractérisés puis appariés [Wu 08, Agarwal 11], comme des points SIFT [Crombez 18]. Il est également possible de déterminer une représentation commune, ainsi, le problème de mise en correspondance 2D/3D est transformé en un problème de mise en correspondance 2D/2D.

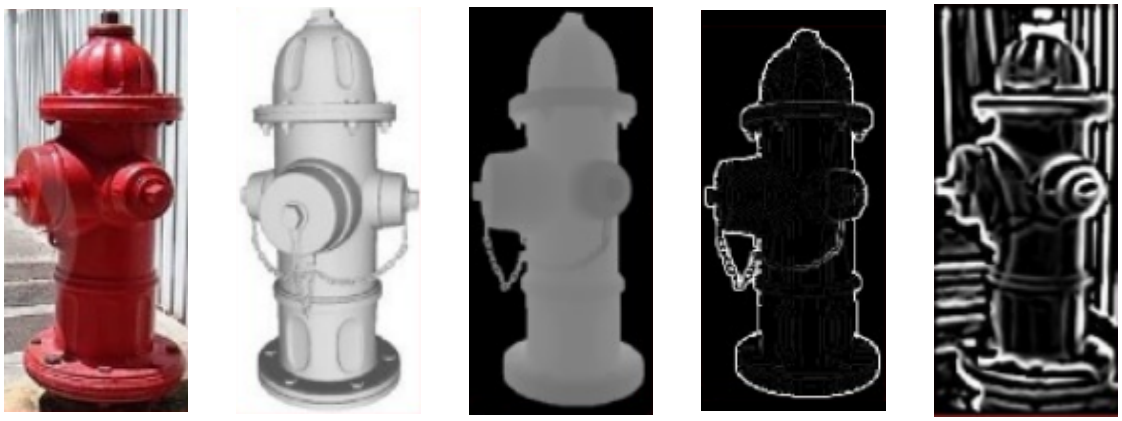
Pour faire correspondre directement des photographies 2D à des modèles 3D ou à des nuages de points, la plupart des systèmes s'appuient sur la détection et la description de

caractéristiques sur les données 2D/3D, puis sur la mise en correspondance de ces caractéristiques. Le début des années 2000 a marqué l'essor du problème de recalage entre une photographie 2D et un modèle 3D d'une scène. De nombreux articles ont alors abordé la question de l'obtention de la pose 3D d'un objet. De manière générale, les approches existantes peuvent être classées en deux catégories : les méthodes indirectes et les méthodes directes [Paudel 14]. Le recalage indirect fait référence aux approches inspirées des processus de recalage 3D-3D ou visant à estimer des paramètres pour un recalage global, telles que l'approche standard Iterative Closest Point (ICP) [Besl 92]. Ces approches sont considérées comme globales, car elles ne fournissent pas d'appariement point à point. En revanche, les approches de recalage direct se concentrent sur la mise en place de correspondances de manière manuelle [Dellepiane 07] ou en exploitant directement des points caractéristiques tels que les points SIFT obtenus en 2D et en 3D [Wu 08, Meierhold 10, Sattler 11, Agarwal 11, Lee 13], ou en utilisant une analyse multi-spectrale [Ricard 05]. Certaines approches privilégient l'utilisation de primitives plus structurées comme des lignes [Kamgar-Parsi 11, Xu 17], des plans [Tamaazousti 11] ou des boîtes englobantes [Liu 05]. Dans le domaine de l'imagerie médicale, par exemple, les contours des objets sont souvent exploités [Cyr 00]. Certaines approches se fondent sur les courbes qui possèdent des caractéristiques géométriques particulières. Par exemple, dans l'étude de [Ohtake 04], les courbes sont supposées être des paraboles, appelées lignes paraboliques, et doivent être situées sur des points de courbure extrême. Cette approche permet de tenir compte des propriétés caractéristiques de la surface de l'objet. Cependant, ces critères ne restent pas invariants en cas de changement de point de vue.

La difficulté principale réside dans la nature des données manipulées : nous n'avons qu'une seule image et aucune contrainte sur les objets manipulés. En effet, une image 2D présente une texture, un fond et un éclairage que nous ne retrouvons pas dans le modèle 3D non texturé.

Plus précisément, dans des images, l'apparence de l'objet dépend, d'une part, des caractéristiques intrinsèques de l'objet, comme la texture, la couleur, l'albédo, et d'autre part, des conditions d'acquisition, comme la pose et l'éclairage. De plus, dans une image prise dans un environnement non contrôlé, le fond (le reste de la scène qui contient l'objet) peut également posséder une texture ou des formes qui ajoutent des difficultés supplémentaires pour comparer le modèle 3D et l'image. À l'inverse, pour un modèle 3D, dans la plupart des cas, la texture de l'objet n'est pas modélisée et il n'y a pas de fond. Nous voulons donc être capables de détecter des primitives d'intérêt, sous la forme de point, qui sont répétables dans les deux modalités. Cependant, dans l'ensemble des travaux cités ci-dessus, la répétabilité des primitives obtenues n'est généralement pas étudiée.

Une autre possibilité, largement utilisée dans la littérature, est de considérer une représentation commune. Nous pouvons envisager deux types de rendu possibles : les images de synthèse [Campbell 01, Rosenhahn 04] ou les cartes de profondeur [Bo 11, Darom 12,



(a) Image. (b) Modèle 3D. (c) Rendu. (d) Carte CS. (e) Carte MCS.

FIGURE II.13. – Détection de points d'intérêt 2D/3D par saillance curviligne multi-échelle [Rashwan 19]. À partir d'une image, (a), et de plusieurs modèles 3D, (b), les auteurs et autrices génèrent une collection de cartes de profondeur obtenues suivant différents points de vue, (c). Ensuite, des primitives d'intérêt sont détectées dans les deux modalités en utilisant la saillance curviligne. Il est alors possible d'obtenir une carte de saillance curviligne CS, (d), pour chaque carte de profondeur et une carte de saillance curviligne multi-échelle MCS, (e), pour chaque image.

Choy 15]. Plus récemment, l'utilisation d'un gradient moyenné, *average shading gradients*, a été proposée. Cette technique consiste à calculer la moyenne des normales des gradients en supposant un ensemble d'éclairages possibles pour tenir compte de l'incertitude sur la source d'éclairage.

Parmi ces trois manières de représenter un objet 3D en 2D, nous avons opté pour les cartes de profondeur. Les cartes de profondeur utilisées sont simplement générées en effectuant différentes projections, suivant différents points de vue, du maillage. Nous pensons que ce type de représentation permet de mieux saisir les formes significatives de l'objet, sans recourir à un rendu synthétique des informations de couleur ou de texture (qui ne sont pas disponibles dans les modèles manipulés). La difficulté de cette représentation réside dans le choix de l'ensemble des points de vue pour générer les cartes de profondeur.

Enfin, pour détecter les points saillants sur nos deux modalités, il nous semble important de favoriser un détecteur qui soit le plus répétable possible. En effet, la comparaison n'aura de sens que si nous comparons les mêmes points de l'objet.

Après avoir sélectionné une représentation commune, il est essentiel d'effectuer la mise en correspondance des primitives entre deux modalités différentes dans cette représentation commune. En général, cette mise en correspondance peut être réalisée de manière partielle ou dense [Crombez 18], locale ou globale [Scharstein 02]. Les primitives détectées peuvent varier en complexité, allant des points d'intérêt aux contours [Lee 07], en passant par les segments et les courbes [Judd 07]. Il s'agit de l'ensemble de courbes dont les

points correspondent à des maxima locaux de la surface. Enfin, certaines approches combinent différents types de primitives, comme cela est décrit dans [Bo 11]. Finalement, dans [Plötz 17], les coins sont détectés en 3D en utilisant le détecteur de Harris en 3D. Pour caractériser les points, les auteurs utilisent les images de gradients moyens obtenus sur les images de rendu. Ainsi, pour chaque image requête, de manière similaire, ils détectent des coins, en multi-échelle, et caractérisent le point en prenant les gradients calculés dans le voisinage. Ainsi, chaque point 2D est apparié avec tous les points de la base de données, en utilisant le descripteur HoG. Le détecteur ontroduit dans [Rashwan 19] s'appuie sur la notion de courbure tout en diminuant l'effet de la texture (grâce à une analyse multi-échelle et la prise en compte du concept de carte de *focus*). Plus précisément, il utilise la saillance curviligne CS, *Curvilinear Saliency*, qui exploite les deux valeurs de la matrice Hessienne, κ_1 et κ_2, comme le détecteur PCBR, mais en calculant : $\kappa_1 - \kappa_2$, avec $\kappa_1 \geq \kappa_2$ et en ajoutant une analyse multi-échelle, d'où le terme de saillance curviligne multi-échelle MCS, *Multiscale Curvilinear Saliency*, cf. Figure II.13. Les résultats présentés dans [Rashwan 19] montrent que ce détecteur permet la meilleure répétabilité des points extraits sur les différentes modalités. De plus, il permet de discriminer les points d'intérêt qui appartiennent à la forme de ceux qui appartiennent à une texture et est capable de faire ressortir les principaux éléments, ceux qui sont les plus saillants dans une donnée. Ce détecteur de contour peut également être utilisé en tant que pré-traitement des données avant de les fournir à un réseau de neurones [Abdulwahab 19]. Ce dernier a pour objectif de faire correspondre une image à un modèle géométrique 3D représenté par un ensemble de cartes de profondeur.

2.7. Bilan et choix

Pour réaliser l'estimation de la qualité de la mise en valeur d'un objet donné par une image en utilisant une extraction de l'information essentielle, nous souhaitons avoir une détection qui soit répétable entre l'image et la carte de profondeur de l'objet obtenu suivant le même point de vue. Intuitivement, nous pensons qu'entre ces deux types de représentation, l'élément commun est la forme de l'objet que nous avons associée à la notion de courbure. Nous voulons à la fois considérer les contours des silhouettes des objets et les contours significatifs à l'intérieur de l'objet d'intérêt, formant l'information essentielle de l'objet. Nous avons cité de nombreux détecteurs en 2D et en 3D, mais nous faisons le choix de nous appuyer sur l'extraction de formes curvilignes, et plus précisément sur l'extraction de la saillance curviligne proposée au sein de mon équipe de recherche par les auteurs et autrices de [Rashwan 19].

En résumé

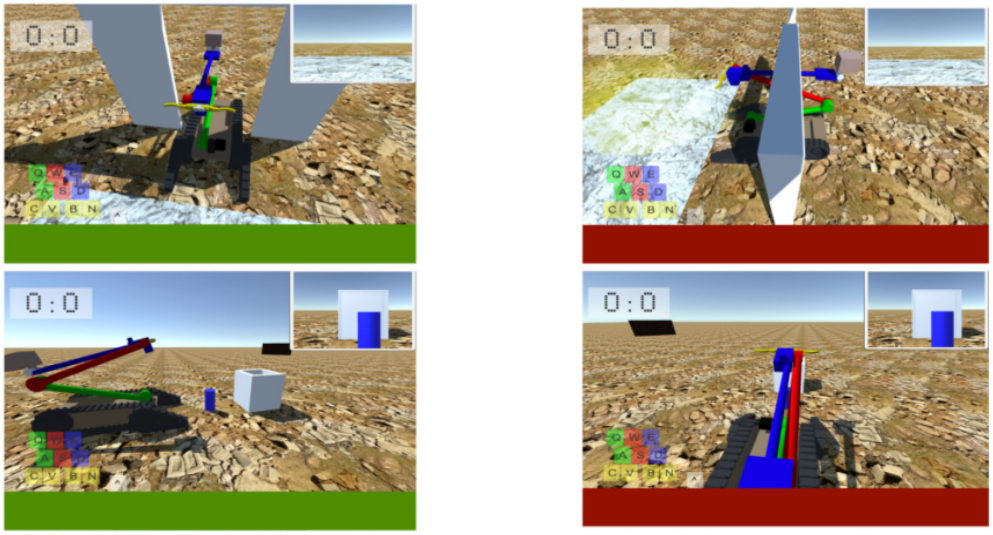
Nous souhaitons estimer la pertinence d'images par rapport à un objet 3D et classer ces dernières en fonction de la mise en valeur de l'objet étudié. Évaluer la qualité des images est une tâche subjective, largement dépendante de la perception individuelle, mais cruciale pour de nombreuses applications. Ainsi, diverses mesures ont été développées pour reproduire au mieux le jugement humain et se rapprocher de la perception visuelle humaine, en considérant principalement le confort visuel et la qualité subjective des images.

Cependant, les méthodes existantes sont limitées par plusieurs aspects. Elles se concentrent généralement sur des caractéristiques globales de l'image, ne prenant pas en compte la pertinence spécifique par rapport à un objet donné. De plus, ces approches traitent souvent des images fortement corrélées ou s'appuyant uniquement sur les caractéristiques visuelles de l'image, sans tenir compte de la géométrie de l'objet représenté. Enfin, la majorité des méthodes cherchent à évaluer subjectivement la qualité des images afin de correspondre au jugement humain, tandis que notre approche se concentre sur l'extraction de l'information essentielle appartenant à l'objet et disponible dans l'image.

Nous avons le souhait de proposer une approche entièrement automatique pour sélectionner et classer les images en fonction de leur capacité à mettre en valeur un objet donné. Contrairement aux approches existantes, nous voulons prendre en compte la géométrie de l'objet 3D représenté dans les images, ce qui lui permet d'extraire des informations pertinentes indépendamment de la texture, de la couleur et de l'éclairage. Notre approche évalue l'intérêt d'une vue 2D d'un objet par rapport à son modèle 3D, en se basant sur la géométrie de cette vue, plutôt que sur la qualité intrinsèque de l'image elle-même. Pour atteindre cet objectif, nous pensons nécessaire d'extraire les éléments saillants d'un objet, en 2D et en 3D, c'est pourquoi dans ce chapitre, nous avons également proposé une revue des techniques existantes avec une attention particulière à celles qui permettent de détecter les points répétables entre les images 2D et les modèles 3D.

3. Sélection de la meilleure vue en 3D

Une caractéristique des modèles 3D est leur indépendance par rapport au point de vue, contrairement aux images. Par conséquent, la sélection du meilleur point de vue à partir d'un modèle 3D offre une variété illimitée de perspectives, contrairement aux images 2D qui sont restreintes par leur nombre disponible initialement. En général, l'objet est présenté hors contexte, autrement dit, la scène dans laquelle le maillage 3D se trouve est dépourvue de tout autre objet environnant. Il n'y a donc aucun risque d'occultation par d'autres objets, ni de variations de lumière ou de couleur.

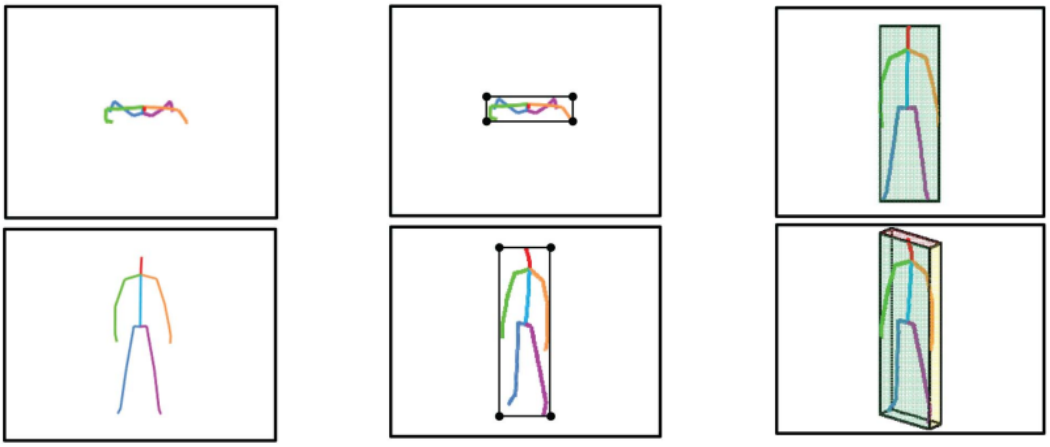


(a) Points de vue optimaux.

(b) Points de vue non optimaux.

FIGURE II.14. – Comparaisons entre un point de vue optimal et un mauvais point de vue pour deux opérations robotiques comme présentées et illustrées dans [Dufek 21]. Sur la première ligne, la tâche à réaliser par le robot est de traverser une porte, tandis que sur la deuxième ligne, l'objectif est de faire manipuler des objets par le robot.

Dans la littérature, diverses approches ont été élaborées afin de choisir les points de vue optimaux d'un maillage 3D, cette tâche étant cruciale dans de nombreux domaines, de la réalité virtuelle à la conception assistée par ordinateur. En effet, choisir la perspective la plus pertinente peut grandement influencer la compréhension et l'interaction avec l'objet représenté. Dans le cas du traitement direct du maillage 3D, la problématique de cette thèse s'apparente au problème de la sélection du meilleur point de vue ou *Best View Selection*. Nous faisons le choix de manipuler des objets 3D non texturés dans diverses catégories, comme les êtres telles que celles des humains, les animaux, les créatures imaginaires, les objets du quotidien, les pièces mécaniques et les objets non familiers, comme les molécules. L'objectif de la sélection de la meilleure vue est d'optimiser et d'améliorer l'expérience visuelle en identifiant et en proposant le point de vue le plus adapté à une



(a) Métrique n°1. (b) Métrique n°2. (c) Métrique n°3.

FIGURE II.15. – Illustrations des trois métriques utilisées dans [Kwon 20] pour déterminer la meilleure vue d'un squelette. Dans les trois paires proposées, les vues de la première ligne sont considérées comme étant de mauvaise qualité, alors que celles de la deuxième ligne sont plus pertinentes pour visualiser le squelette. La première métrique, (a), est associée à la longueur des membres du squelette, la deuxième, (b), concerne la surface visible de la boîte englobante en 2D tandis que la dernière, (c), est liée à la surface visible d'une boîte englobante 3D.

tâche, une application ou une interaction avec des utilisateurs et utilisatrices. La génération automatisée d'un point de vue optimal est primordiale pour une variété de tâches comme la simplification de maillage [Lee 05], la reconnaissance d'objets [Vázquez 01], la compréhension de scène dynamique [Biswas 23], la visualisation de réduction dimensionnelle [Castelein 23], la simulation de phénomènes physiques [Lee 11], l'optimisation de l'éclairage pour le rendu [Naraoka 07], la réalisation d'opérations robotiques [Dufek 21], comme celles illustrées dans la Figure II.14, et dans le cadre de chirurgies mini-invasives assistées par robots [Su 21].

Sélectionner le point de vue idéal peut concerner diverses données, telles que des maillages 3D, des nuages de points en 3D, des squelettes ou des données volumiques. Par exemple, l'approche proposée dans [Tao 16], utilise un modèle de sélection de points de vue s'appuyant sur la similarité entre images pour apprendre comment les experts et expertes en visualisation choisissent des points de vue représentatifs de volumes. Plus précisément, plusieurs rendus des volumes, selon différents points de vue, sont générés et reçoivent un score sur la base d'une mesure de similarité, qui évalue la forme spatiale et la similarité d'apparence entre les images sélectionnées par des experts et des expertes et les images de rendu. Le point de vue optimal est celui qui a reçu le score maximal.

Dans un domaine différent, les travaux détaillés dans [Kwon 20] explorent une métrique qui peut être utilisée pour quantifier la vue d'un squelette ou d'un modèle humain en

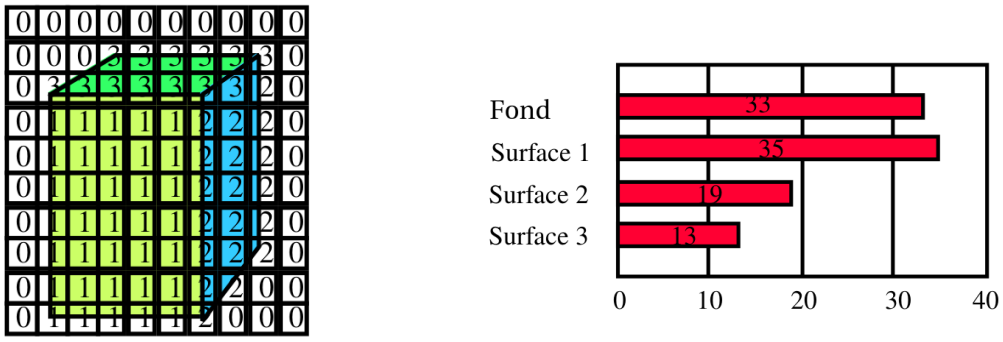
3D, et de déterminer l'angle de caméra le plus favorable, conformément aux évaluations subjectives effectuées par les participants et participantes à leur étude. Plus précisément, ils définissent trois métriques, illustrées dans la Figure II.15, et formulent un problème d'optimisation de point de vue, dont la fonction objective est la somme des trois métriques. La première métrique est liée à la longueur des membres du squelette. Sur la base de leur étude subjectif, ils peuvent affirmer que plus la somme des longueurs des membres est importante, plus l'utilisateur à tendance à percevoir correctement les informations de visualisation. La deuxième métrique concerne la surface visible de la boîte englobante en 2D : la visibilité du squelette augmente lorsque la surface de la boîte englobante 2D est grande. Enfin, la dernière métrique est liée à la surface visible d'une boîte englobante 3D. Lors de l'étude, les utilisateurs et utilisatrices ont eu tendance à préférer l'apparence du squelette proportionnellement au volume 3D.

Dans la suite de cette section, nous avons choisi de présenter les méthodes existantes, des plus anciennes aux plus récentes, en distinguant celles qui font appel à des outils classiques d'extraction d'attributs géométriques et nous terminons pas les techniques les plus récentes basées apprentissage profond. Enfin, le dernier paragraphe nous permettra de faire le bilan de tout ce que nous avons présenté et d'introduire la direction que nous souhaitons prendre.

D'après les travaux menés dans [Bonaventura 18] et [Secord 11], les méthodes de sélection des points de vue peuvent être catégorisées en fonction de la catégorie des attributs qu'elles extraient : surface, silhouette, profondeur, stabilité, courbure et utilisation d'information *a priori*. Dans la suite, nous reprenons chacun de ces éléments.

3.1. Surface

Les mesures détaillées dans cette section s'appuient principalement sur le calcul des surfaces visibles des objets 3D, dans chaque point de vue étudié. Par exemple, les auteurs de [Plemenos 96] ont évalué la qualité du point de vue en utilisant le nombre de triangles visibles, considérant que les régions les plus significatives contiennent plus de détails et donc plus de triangles. Cependant, cette mesure peut être sensible à la discrétisation du modèle 3D. Afin de contourner ce problème, les auteurs ont également pris en compte la surface projetée du modèle 3D comme mesure de sa qualité, dans des travaux ultérieurs [Plemenos 04, Plemenos 06]. Plus précisément, la complexité d'un point de vue a été définie par la somme de deux valeurs : le pourcentage de faces visibles et le pourcentage de pixels occupés par les surfaces 2D projetées de l'objet, cf. Figure II.16. Une face peut être visible selon plusieurs points de vue, elle a donc une influence binaire, alors que sa surface projetée peut occuper un nombre variable de pixels sur l'image 2D de la vue. Ensuite, les auteurs de [Vázquez 01, Vázquez 03, Page 03, Polonsky 05] ont introduit une mesure de sélection des meilleurs points de vue utilisant l'entropie de Shannon [Shannon 48, Yeung 08] sur les



(a) Pixels occupés dans l'image de la vue considérée

(b) Nombre de pixels occupés par surface

FIGURE II.16. – Évaluation de la qualité d'un point de vue 3D basé surface. L'approche présentée dans [Plemenos 04] utilise de manière traditionnelle le pourcentage de faces visibles, tout en introduisant un nouveau terme correspondant au pourcentage de pixels occupés par les surfaces 2D projetées de l'objet, comme illustré visuellement en (a) où chaque surface est numérotée et quantitativement en (b).

différentes surfaces projetées, et considérant ainsi que la quantité d'informations capturées par un point de vue spécifique permet d'estimer sa qualité. Plus précis ément, l'entropie de Shannon est appliquée aux probabilités des faces projetées. La probabilité d'une face correspond au rapport entre la surface projetée visible de cette face et la surface projetée totale visible de l'objet. La mesure proposée dans [Sbert 05] évalue la qualité du point de vue en comparant la distribution normalisée des aires projetées des polygones, à partir d'un point de vue, avec la distribution normalisée des aires réelles des polygones, à l'aide de la distance de Kullback-Leibler.Enfin, dans [Feixas 09], la méthode proposée s'appuie sur l'information mutuelle des points de vue ou *viewpoint mutual information*, qui capture le degré de corrélation entre un point de vue et l'ensemble des polygones. Les valeurs les plus faibles correspondent aux vues les plus représentatives ou les plus pertinentes, autrement dit celles qui montrent le plus grand nombre possible de polygones de manière équilibrée. Cette mesure est conçue pour être insensible à la discrétisation du modèle.

3.2. Silhouette

Les techniques de cette section utilisent la silhouette de l'objet pour évaluer la pertinence de chaque point de vue étudié. Dans [Polonsky 05], les auteurs et autrices ont proposé d'utiliser la longueur en pixels de la silhouette projetée du modèle comme mesure de la qualité du point de vue, en faisant l'hypothèse que le meilleur point de vue est celui qui présente une silhouette de longueur maximale. Par ailleurs, dans [Page 03], c'est l'entropie de la distribution de la courbure de la silhouette qui a été introduite comme mesure de la

qualité. D'autres métriques, s'appuyant également sur la silhouette, ont été développées, comme par exemple, dans [Vieira 09] où l'intégrale totale de la courbure de la silhouette a été définie pour représenter la complexité de la silhouette ou encore [Secord 11] qui s'appuie sur les extrema de la courbure.

3.3. Profondeur

Les auteurs de [Stoev 02] ont remarqué que la surface projetée n'était pas suffisante pour visualiser les terrains, car la vue la plus étendue est généralement celle du dessus. Ils ont donc présenté une méthode de placement des caméras qui optimise la profondeur maximale de l'image, en plus de la surface projetée. Cette profondeur maximale correspond à la profondeur 3D la plus grande sur l'objet, étant donné un point de vue. Ils définissent le meilleur point de vue comme étant celui qui maximise la surface projetée ainsi que la profondeur maximale et minimise la différence entre la surface projetée et la profondeur maximale. Dans [Secord 11], la profondeur maximale manipulée correspond à celle utilisée dans [Stoev 02], et est considérée comme un descripteur de la qualité du point de vue. Une étude plus récente, détaillée dans [Marsaglia 21], a introduit trois nouvelles mesures de la qualité du point de vue capables de prédire les préférences humaines en matière de placement de caméras. Plus précisément, cette étude analyse les possibles combinaisons de ces trois métriques, qui s'appuient toutes sur l'entropie de Shannon. La première estime l'entropie de la profondeur, une autre calcule l'entropie des données 3D et la dernière fait intervenir l'entropie de l'ombrage. L'hypothèse dans cette étude est que les mesures utilisant de l'entropie permettent d'excellentes prédictions, notamment lorsqu'elles sont combinées.

3.4. Notion de stabilité

Les mesures s'appuyant sur la stabilité calculent la corrélation entre un point de vue spécifique et ses voisins. Par exemple, dans [Feixas 09], cité précédemment, l'instabilité d'un point de vue est définie à partir de la notion de dissimilarité entre deux points de vue. Cette dissimilarité est calculée à l'aide de la divergence de [Burbea 82] appliquée sur les distributions d'aires projetées des deux points de vue considérés. De manière similaire, la méthode introduite dans [Vázquez 09] permet de calculer la stabilité de la vue à partir des cartes de profondeur de tous les points de vue. Le degré de similitude entre deux points de vue est donné par la distance de compression normalisée entre deux cartes de profondeur. Deux vues sont considérées comme similaires si leur distance est inférieure à un seuil donné. Ainsi, la vue la plus stable est celle qui présente le plus grand nombre de vues similaires.

3.5. Courbure et notion de saillance

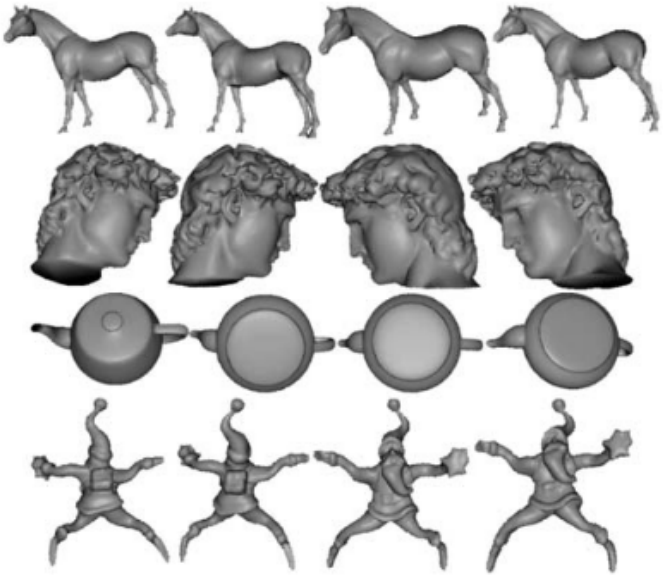


FIGURE II.17. – Classements de vues s'appuyant sur l'entropie des distributions des courbures . L'illustration est extraite de [Polonsky 05]. Il s'agit des quatre meilleures vues, de la meilleure (à gauche) à la quatrième meilleure (à droite).

Les mesures détaillées dans cette section s'appuient sur la courbure des surfaces visibles. Il convient de noter que, dans certaines mesures, d'autres attributs, tels que les attributs de surface, peuvent également être pris en compte. Dans [Polonsky 05], une des mesures proposées évalue l'entropie de la distribution de courbure issue de la partie visible de l'objet. Cette mesure s'inspire de l'entropie de la distribution de courbure gaussienne définie par [Page 03]. Un exemple de sélection de meilleures vues pour divers modèles 3D issus de [Polonsky 05], est présenté dans la Figure II.17. Lorsqu'on utilise la saillance, comme dans [Lee 05, Rudoy 12, Nouri 15, Habibi 15], un score correspondant à la somme des saillances intrinsèques des sommets visibles est associé à chaque point de vue, cf. Figure II.18. Plus précisément, les auteurs de [Lee 05] ont présenté une mesure permettant de sélectionner le meilleur point de vue en fonction de l'importance de la saillance observée. Pour cela, une saillance intrinsèque est calculée pour chaque sommet à l'aide de la courbure définie dans [Taubin 95]. Le score attribué à chaque point de vue étudié correspond à la somme des saillances des sommets visibles. Plus cette mesure est élevée, plus le point de vue étudié est considéré comme intéressant. Malheureusement, cette mesure est sensible à la discrétisation polygonale puisque la somme est effectuée pour les sommets visibles. Pour rendre ce processus de sélection plus rapide, une étape d'optimisation par descente de gradient est appliquée. Ainsi, seul un nombre réduit de points de vue stratégiques sont analysés. Comme dans [Lee 05], les auteurs de [Sokolov 05] ont présenté une mesure de

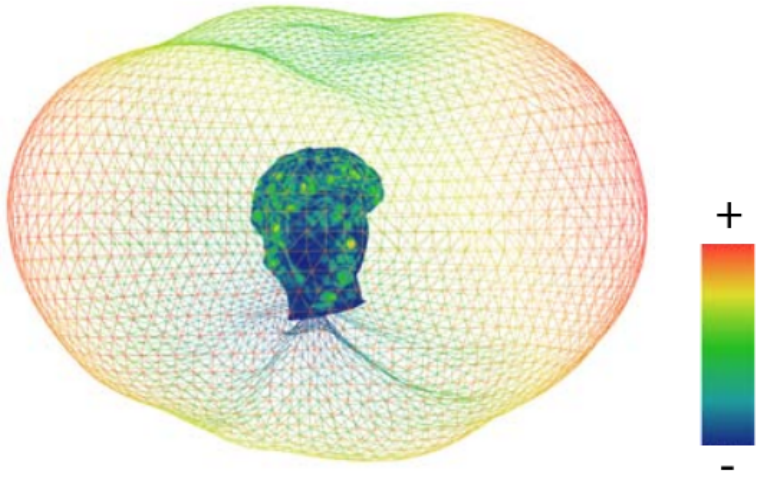


FIGURE II.18. – Pertinence d'un ensemble de vues en fonction de la saillance. L'illustration provient de [Lee 05]. Le meilleur point de vue est celui qui maximise la somme des saillances des points visibles. La couleur du maillage filaire autour du modèle de la tête de David montre l'amplitude de cette somme.

la qualité s'appuyant sur la somme des courbures visibles, pour un point de vue donné. D'autres approches, comme celle introduite dans [Leifman 16], adoptent une idée similaire en incorporant une pondération de chaque sommet exploitant l'angle entre la normale de la surface au sommet et la direction d'observation. Ainsi, les sommets faisant face à la caméra sont privilégiés dans le calcul de la saillance.

De même, en s'inspirant de l'approche présentée dans [Lee 05], les auteurs de [Feixas 09] sélectionnent la meilleure vue en utilisant une mesure s'appuyant sur la saillance des polygones. En disposant un maillage au centre d'une scène fermée, les auteurs ont introduit un modèle théorique s'appuyant sur l'information mutuelle polygonale. Cette approche définit la saillance du maillage en fonction du degré de corrélation entre un polygone et l'ensemble des points de vue. En d'autres termes, un polygone situé au cœur d'une zone lisse aura tendance à avoir une faible saillance, car les polygones environnants présenteront de faibles différences de visibilité par rapport à l'ensemble des points de vue.

Souvent, les méthodes de détection de saillance 3D mentionnent que l'une des applications possibles à l'approche proposée pourrait être la sélection de la meilleure vue, comme l'illustre l'étude détaillée dans [Limper 16].

3.6. Utilisation d'information *a priori*

Les auteurs de [Becker 07] analysent comment les singularités intrinsèques à l'objet, qui attirent l'attention de l'observateur, peuvent être détectées en considérant les connaissances *a priori* sur les objets manipulés. Par exemple, lorsque le modèle 3D est une créature

avec des yeux ou un visage, il a été observé dans [Zusne 70] que les individus préfèrent les vues où les yeux sont visibles. Sur un autre aspect, les auteurs de [Podolak 06] ont introduit une méthode pour choisir automatiquement de bons points de vue en minimisant la symétrie des objets observés depuis un point de vue. Pour prendre en compte la sémantique dans la sélection de la meilleure vue, les autrices de [Mortara 09] utilisent une segmentation basée sémantique pour extraire automatiquement les caractéristiques significatives d'une forme 3D. En particulier, elles proposent une nouvelle fonction de classement des points de vue qui combine les critères de visibilité avec un raisonnement sur la proportion de caractéristiques pertinentes visibles, extraites à partir de la segmentation.

3.7. Évaluation et combinaison des attributs géométriques

Certains travaux ont choisi d'étudier tous les attributs connus pour sélectionner le meilleur point de vue et d'évaluer leurs performances. Certains vont même jusqu'à proposer des combinaisons d'attributs pour optimiser les résultats [Secord 11]. Après avoir analysé sept attributs géométriques différents, les auteurs et autrices de [Polonsky 05] ont conclu qu'aucun n'était parfait et ont suggéré qu'une combinaison de ces attributs amplifierait leurs avantages respectifs, les uns par rapport aux autres. Leurs expériences suggèrent également que pour accélérer les calculs, il est avantageux de discrétiser les vues en s'appuyant sur les résultats de [Blanz 99], qui notent que les vues de trois quarts des objets sont souvent les plus représentatives. Dans l'étude comparative menée dans [Dutagaci 10], un ensemble de sept méthodes de sélection de la meilleure vue, issues de l'état de l'art, a été étudié. Pour vérifier leur corrélation avec les préférences des êtres humains, une étude utilisateurs et utilisatrices a été réalisée. Une manière pour quantifier l'erreur entre les vues, déterminées par les méthodes de l'état de l'art, et les vues sélectionnées au cours de l'étude, a également été développée. D'après leurs résultats, les deux meilleurs attributs à utiliser correspondent à celui de la saillance intrinsèque de l'objet et celui s'appuyant sur l'entropie des probabilités des faces projetées. Pour rappel, la probabilité d'une face correspond au rapport entre la surface projetée visible de cette face et la surface projetée totale visible de l'objet. Le modèle présenté dans [Secord 11] pour sélectionner la meilleure vue d'un objet 3D correspond à un modèle linéaire. En effet, l'objectif est de déterminer une combinaison linéaire d'attributs géométriques qui imite les préférences humaines, récoltées au sein d'une étude par des utilisateurs et des utilisatrices, en matière de pertinence de la vue. Quatorze critères organisés en cinq catégories liées à différents aspects d'une vue, tels que la surface ou la silhouette, ont été étudiés. D'après les résultats, les deux meilleurs attributs correspondent à la quantité de surface visible ainsi que l'entropie des surfaces des faces projetées. Les auteurs soulignent également l'importance de la prise en compte des yeux dans le choix des points de vue par les utilisateurs et utilisatrices, lorsque que la détection des yeux est possible. Enfin, plus récemment, les auteurs de [Bonaventura 18]

ont examiné un ensemble de vingt-deux mesures pour la sélection des points de vue, dont onze n'avaient pas été examinées auparavant. Ils ont notamment étendu la classification réalisée par [Secord 11]. La base de données issue de l'étude des utilisateurs et utilisatrices, réalisée dans [Dutagaci 10], a été utilisée comme vérité terrain. D'après leurs résultats, les deux meilleurs attributs à utiliser correspondent à celui qui comptabilise le nombre de faces visibles et celui correspondant à l'entropie de la courbure visible.

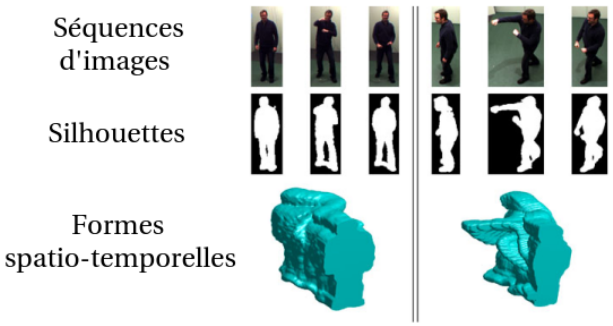


FIGURE II.19. – Illustrations de formes spatio-temporelles associées aux séquences d'images. Cette figure est extraite de [Rudoy 12].

La combinaison d'attributs caractéristiques a aussi été utilisée dans d'autres domaines, tels que la sélection de la meilleure vidéo d'une scène dynamique. Par exemple, les auteurs de [Rudoy 12] évaluent automatiquement la qualité du point de vue offert par une vidéo parmi un ensemble de vidéos filmant la même scène dynamique. Pour déterminer les meilleurs points de vue, les auteurs proposent de calculer un score de visibilité. Ce score s'appuie sur la combinaison entre des caractéristiques temporelles, des attribiuts liés à l'espace occupé par la personne filmée, et d'autres propriétés issues de la forme spatio-temporelle induite par les silhouettes de l'acteur ou de l'actrice au sein de chaque vidéo, comme illustré dans la Figure II.19.

3.8. Approches basées apprentissage

Dans cette section, nous abordons les techniques à base de classifieurs avant d'aborder les approches plus récentes qui s'appuient sur un apprentissage profond.

Dans [Laga 10], l'approche repose sur un principe courant en classification, c'est-à-dire que les modèles d'une même classe ont des caractéristiques distinctives similaires, qui les séparent des modèles des autres classes. À partir de cette observation, le classifieur proposé apprend à identifier les vues 2D qui maximisent la similitude entre les objets de la même classe. Dans les travaux détaillés dans [Liu 12], et qui sont proches de ceux présentés dans [Hall 05, Mezuman 12], pour un objet 3D donné, un ensemble d'images est collecté sur internet, à partir duquel on extrait les points de vue présents dans les images. Une

étude statistique est réalisée pour déterminer la fréquence d'apparition de chaque point de vue. Le plus fréquent est considéré comme celui offrant la meilleure vue, cf. Figure II.20. L'originalité des travaux de [Zhao 15] réside dans le fait de considérer qu'un point de vue d'un objet 3D est pertinent si un humain le dessine habituellement à partir de ce même point de vue, cf. Figure II.21.

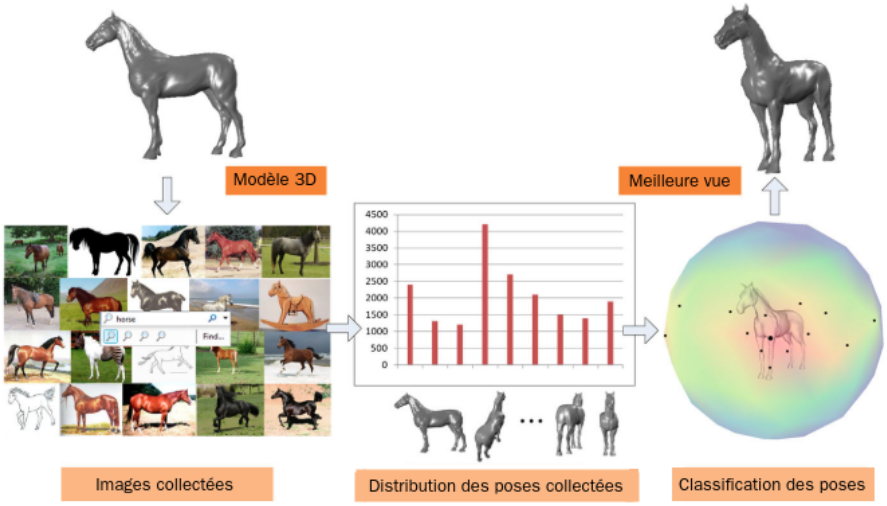


FIGURE II.20. – Sélection du meilleur point de vue à partir d'une étude statistique de collection d'images récupérées sur internet. Cette illustration est extraite de [Liu 12]. Les meilleurs points de vue sont ceux qui sont les plus fréquents dans la collection.

L'approche de [Kim 17] repose sur deux architectures de réseaux de neurones convolutifs, *Convolutional Neuron Network*, CNN, qui encodent des informations spécifiques de chaque catégorie, apprises à partir d'un ensemble de formes 3D et d'images 2D récoltées sur internet. Le premier CNN est utilisé pour déterminer l'orientation verticale naturelle des objets, tandis que le deuxième CNN évalue les projections et les vues saillantes après alignement vertical, en exploitant les informations de saillance extraites dans les images et dont on suppose qu'elles correspondent aux préférences humaines, mais également la saillance des objets. Le travail présenté dans [Yang 19] est similaire et la contribution principale réside dans l'utilisation d'une structure résiduelle de plusieurs niveaux. Nous pouvons considérer l'approche de [Hartwig 22] comme assez proche également mais, dans cette publication, c'est une architecture de réseau siamois [5] qui est utilisée pour prédire les points de vue préférés des êtres humains à partir d'un modèle 3D. Les travaux de [Schelling 21] se dintinguent des travaux que nous avons déjà mentionnés par le fait qu'ils introduisent un processus de génération dynamique d'étiquettes. Ce processus adapte les

5. Un réseau siamois est une architecture de réseau de neurones utilisée pour la comparaison de paires d'entrées. Ils sont constitués de deux branches symétriques partageant les mêmes poids. Chaque branche traite une des deux entrées, puis leurs représentations sont comparées pour mesurer leur similarité.

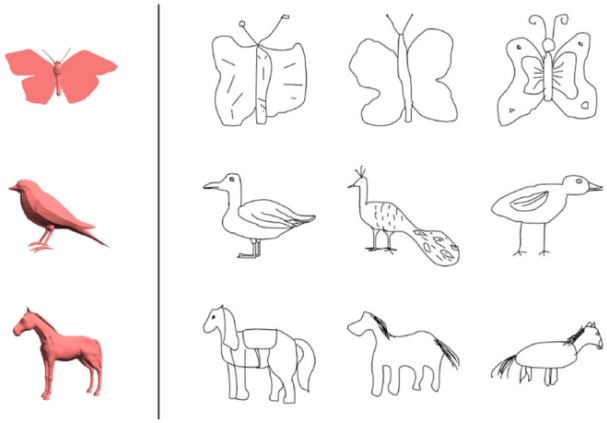


FIGURE II.21. – Sélection de meilleurs points de vue utilisant des croquis humains. Ces travaux s'appuient sur l'analyse des points de vue préférés des utilisateurs et utilisatrices à partir de leurs croquis [Zhao 15]. À gauche différents modèles 3D positionnés selon les vues extraites des croquis fréquemment dessinés, présentés à droite.

sorties du modèle pour résoudre les ambiguïtés des étiquettes, rendant ainsi la décision d'étiquetage adaptative aux prédictions actuelles du réseau. Les ambiguïtés surviennent souvent en raison de la non unicité du meilleur point de vue, notamment avec les objets symétriques. De plus, cette approche a réussi à réduire l'influence de la qualité de subdivision du maillage en prédisant les points de vue à partir du nuage de points non structuré plutôt que du maillage 3D polygonal. Enfin, de manière classique et comme évoqué en 2D, nous retrouvons des approches exploitant la notion d'esthétisme comme celle de [Zhang 20]. L'approche intègre plusieurs facteurs esthétiques, comme l'utilisation de la règle des tiers en photographie, ainsi que des facteurs géométriques (silhouette, courbure).

3.9. Analyses des approches de sélection de la meilleure vue 3D

Dans ces travaux de thèse, nous avons pris la décision de ne pas utiliser une approche basée apprentissage. Nous avons privilégié des méthodes classiques s'appuyant sur l'extraction d'attributs géométriques disponibles directement ou facilement calculables à partir du maillage 3D. Nous avons réalisé ce choix car, l'utilisation de méthodes d'apprentissage nécessiterait un ensemble de données volumineux et diversifié pour l'entraînement et nous n'avions pas accès à ce type de base de données dans la littérature. De plus, construire cette base de données n'était pas envisageable dans le contexte de cette thèse. De plus, nous souhaitions mettre en œuvre une approche directement compréhensible car l'objectif de cette thèse ne concernait pas l'explicabilité des modèles d'apprentissage. En identifiant et en évaluant des caractéristiques géométriques spécifiques de l'objet 3D, telles que le rapport des faces visibles, la saillance des sommets ou la courbure, nous pouvons obtenir des mesures objectives et explicables de la qualité d'un point de vue. De plus, ces méthodes

géométriques sont souvent plus simples à mettre en œuvre et moins coûteuses en temps de calculs par rapport aux approches basées apprentissage profond.

Après avoir examiné attentivement l'état de l'art, nous avons opté pour une approche hybride, exploitant les avantages des méthodes s'appuyant sur l'entropie et celles utilisant une détection de saillance. Plus précisément, nous avons souhaité concevoir une approche s'appuyant sur l'extraction d'attributs géométriques, tels que ceux liés à la surface et la courbure qui offre des avantages tels que la simplicité, la rapidité et l'interprétabilité.

Inspirées par les travaux décrits dans [Polonsky 05, Secord 11, Marsaglia 21], qui ont démontré l'efficacité de l'utilisation conjointe de plusieurs attributs géométriques dans la sélection des meilleures vues, nous avons décidé de combiner plusieurs attributs. Suite aux études comparatives réalisées dans [Dutagaci 10, Secord 11, Bonaventura 18], nous avons sélectionné trois aspects majeurs de la géométrie de l'objet : la proportion de surfaces visibles par rapport à la surface globale, la proportion de la surface des yeux visibles par rapport à la surface globale et enfin la saillance intrinsèque du maillage.

L'importance des yeux a été mentionnée dans divers travaux, nous avons donc inclus un terme correspondant, apportant ainsi une dimension supplémentaire à l'évaluation des points de vue.

Enfin, rappelons que notre objectif est d'évaluer et de déterminer automatiquement les différentes vues d'un objet 3D en fonction de leur pertinence. Notre approche vise à identifier et à retenir exclusivement les représentations 2D qui capturent l'information la plus significative de l'objet 3D, que nous appelons l'information essentielle. En intégrant la saillance intrinsèque du maillage 3D, nous favorisons ainsi les éléments visibles les plus pertinents dans chaque vue, formant ainsi l'information essentielle disponible.

4. Détection de saillance 3D

Lorsque nous avons présenté la façon d'extraire les points de vue les plus pertinents en 3D, nous avons évoqué la notion de saillance. Dans cette dernière section, nous reprenons cet aspect en présentant les approches les plus significatives.

La saillance des sommets peut être utilisée pour mettre en évidence les zones d'intérêt et diriger l'attention visuelle vers les caractéristiques les plus importantes d'un modèle 3D. Ces caractéristiques peuvent être d'ordre géométrique, se distinguant par leur structure particulière, ou sémantique, c'est-à-dire ayant une signification pour les êtres humains. De nombreuses applications graphiques ont utilisé la saillance des maillages pour adapter le traitement d'un contenu 3D. Les auteurs de [Gu 14] ont combiné des techniques de détection de saillance avec l'échantillonnage de Poisson, pour la compression adaptative des images de profondeur. Nous pouvons également citer des applications comme la simplification de maillages [Lee 05], l'échantillonnage de maillages [Wu 13], la mise en correspondance de formes [Gal 06], la reconstruction de surfaces [Song 12a, Song 14] et la

modélisation de foules [McDonnell 09].

Dans la suite, nous avons fait le choix de classer les méthodes de la façon suivante : les approches utilisant une information locale, celles exploitant une information globale, celles s'appuyant sur la notion d'entropie, les méthodes faisant intervenir une technique de classification et enfin les plus récentes à base d'apprentissage.

4.1. Utilisation d'une caractérisation locale

Dans [Miao 10], les auteurs ont proposé une méthode pour estimer la saillance en fonction de la hauteur du relief à partir du maillage 3D. Autrement dit, ils essayent d'adapter une surface à chaque sommet, à différentes échelles, et les écarts les plus importants signifient une plus grande saillance. De plus, à partir de cette saillance par sommet, ils proposent un processus d'extraction des lignes de crête et de vallée le long des directions des courbures principales.

Certaines méthodes utilisent des techniques multi-échelles pour calculer la saillance des sommets car, dans la littérature, il est habituel de considérer que des niveaux d'échelle faible, permettent la mise en évidence de détails fins et singuliers, tandis que des niveaux d'échelle plus élevé, des informations plus globales et générales. Par exemple, dans [Lee 05], les auteurs ont estimé la saillance des maillages 3D à l'aide de filtres gaussiens appliqués sur les courbures moyennes des sommets, à différentes échelles. La saillance finale est obtenue en combinant les saillances des différentes échelles avec une normalisation non linéaire.

Il existe également des méthodes de détection de saillance s'appuyant sur les attributs spectraux du maillage d'entrée, notamment le spectre log-laplacien, comme dans [Song 14]. En effet, les auteurs considèrent que l'irrégularité de ce spectre est étroitement liée à la saillance du maillage, et utilise cette information pour obtenir un résultat de détection de la saillance plus précis en appliquant une analyse multi-échelle dans le domaine spatial.

Dans [Leifman 16], une fraction de sommets présentant des valeurs de distinction élevées en tant que points d'attention sont détectés. Puis, les distances géodésiques entre chaque sommet et les points d'attention les plus proches, ce qui introduit la notion de région d'intérêt, ou de *patches*, sont calculées, cf. Figure II.22. À travers la notion de patches, les auteurs ont voulu modéliser l'idée que les formes visuelles peuvent posséder un ou plusieurs centres de gravité autour desquels la forme est organisée. Par conséquent, les régions proches de ces centres d'attention devraient être plus intéressantes que les régions éloignées. L'algorithme introduit dans [Jeong 17] permet la détection de la saillance 3D indépendamment de la vue, pour les maillages semi-réguliers. Plus précisément, pour chaque échelle, le maillage est représenté sous forme d'un graphe orienté entièrement connecté, où les faces servent de nœuds. Les poids des arêtes sont calculés à l'aide des caractéristiques de courbure locale, et une marche aléatoire ou *random walk*[6], est effectuée pour

6. Une marche aléatoire est une modélisation mathématique du mouvement aléatoire sur un graphe,

(a) Distinction des sommets. (b) Extrémités du modèle 3D. (c) Association de patch

FIGURE II.22. – Caractérisation locale par distinction des sommets, position des extrémités et association des patches. Cette illustration est extraite de [Leifman 16] et permet de présenter les trois concepts utilisés pour déterminer la meilleure vue d'un objet 3D. En (a), un sommet est distinct des autres si son descripteur est différent de ceux de ses voisins. En (b), les extrémités du modèle 3D sont considérées et en (c), il s'agit des régions proches de zones d'attention.

obtenir la distribution des fréquences de visite de chaque nœud, sous la forme d'une carte de saillance. Ensuite, seules les valeurs maximales de saillance sont conservées lors de la fusion des différentes distributions issues des différentes échelles.

4.2. Utilisation d'une caractérisation globale

Les méthodes s'appuyant sur la courbure utilisent une information plutôt locale et cela à tendance à identifier à tort des régions bosselées ou simplement bruitées comme étant saillantes. Les approches s'appuyant sur le calcul du contraste globale visent à diminuer ce défaut. Ainsi, les auteurs de [Sipiran 13] ont mis au point une méthode globale utilisant la méthode Harris 3D, proposée dans leurs précédents travaux [Sipiran 11], correspondant à une extension du célèbre détecteur de Harris. La réponse du détecteur 3D par sommet est alors assimilée à une valeur de saillance. Dans [Wang 15], un descripteur de forme a été introduit afin de contenir une quantité suffisante d'informations pour décrire la structure globale du maillage, ainsi qu'une grande quantité d'informations sur les formes répétées et redondantes, qui peuvent être considérées comme non saillantes. À partir de ces descripteurs, une matrice, dite de structure globale, est extraite et on lui applique une décomposition de faible rang, puis une analyse éparse afin d'obtenir la saillance du maillage, relative à sa structure globale. Enfin, dans [Arvanitis 20], la méthode introduite exploite les informations globales par le biais d'une analyse des composantes principales

qui a été largement utilisée pour détecter la saillance des images.

pour prédire la saillance 3D sur des objets 3D industriels. Ces méthodes visent à supprimer les motifs répétés sur la surface d'un maillage en étendant la comparaison des régions d'un contexte local à un contexte plus large. Toutefois, elles présentent certaines limitations, notamment la nécessité d'une segmentation du maillage et le calcul de distances géodésiques, qui peuvent être problématiques en présence de défauts topologiques tels que les trous et les structures non-manifold.

4.3. Approches basées entropie

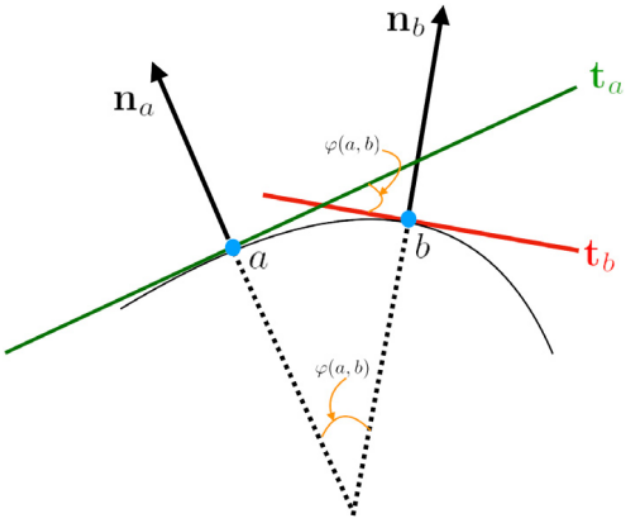


FIGURE II.23. – Visualisation de l'angle tropical utilisé pour la saillance 3D. L'angle tropical $\varphi(a,b)$ entre deux sommets a et b, utilisé dans [dos Anjos 23], correspond à la différence angulaire entre les deux plans tangents t_a en a et t_b en b, définis par les vecteurs normaux n_a et n_b, respectivement. La saillance associé à un sommet v est définie comme la valeur maximale de l'angle tropical entre n'importe quelle paire de sommets dans le voisinage direct de v.

Les méthodes de détection de saillance 3D utilisant l'entropie offrent une nouvelle perspective pour évaluer la pertinence des différentes régions d'une surface en fonction de la quantité d'information qu'elles contiennent. En effet, l'entropie mesure le niveau d'incertitude ou de désordre dans un système. Dans le contexte de la saillance 3D, l'entropie peut être utilisée pour évaluer la distribution des caractéristiques saillantes sur un maillage. Une distribution non uniforme et bruitée aura une entropie plus élevée, ce qui peut être interprété comme une saillance plus forte. Les auteurs de [Page 03] proposent une méthode d'analyse des formes qui définit la saillance à partir de l'entropie de Shannon appliquée sur les courbures gaussiennes de la surface du maillage. L'approche introduite dans [Limper 16] permet de réaliser une estimation de la saillance des modèles 3D en s'appuyant sur

l'entropie de courbure locale. Cette méthode permet de classer les régions d'une surface 3D en fonction de la quantité d'informations qu'elles contiennent. En appliquant l'entropie de Shannon aux données de maillage 3D, la méthode détermine efficacement la saillance du maillage en capturant différents types de caractéristiques saillantes à plusieurs échelles et en les combinant en une seule carte de saillance.Enfin, la méthode proposée dans [dos Anjos 23] est une méthode multi-échelle permettant de capturer les régions et les plis saillants locaux et globaux d'un maillage. Cette approche repose sur l'entropie de l'angle tropical de courbure qui estime la différence maximale des normales voisines pour chaque sommet d'une surface, cf. Figure II.23. Avec cette technique, un calcul robuste de la saillance des maillages est garantie, même pour les modèles denses comportant des millions de polygones.

4.4. Approches basées classification

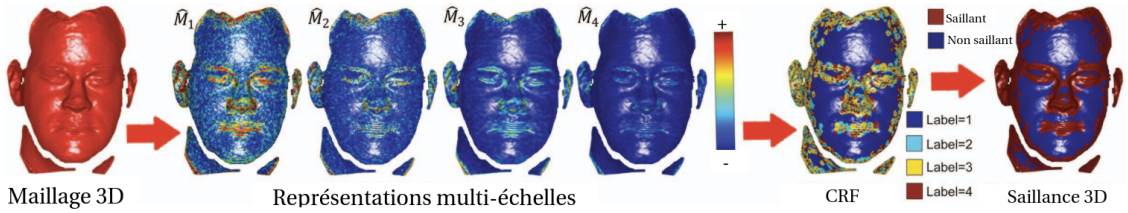


FIGURE II.24. – Estimation de la saillance 3D à l'aide de champs aléatoires conditionnels. L'illustration est extraite de [Song 12b]. L'approche consiste à définir une représentation multi-échelle de la surface qui permettra ensuite de réaliser une classification des différents points de la surface afin d'en extraire les points saillants, comme illustré à droite.

Les approches globales mentionnées précédemment, peuvent nécessiter des calculs plus intensifs. Pour éviter ces inconvénients, certaines méthodes effectuent une étape de classification pour estimer de la saillance 3D. Par exemple, les auteurs de [Gal 06] proposent de classifier un maillage 3D, de manière éparse, à partir de ses propriétés géométriques, et de représenter chaque classe par un descripteur. Ce dernier contient des informations liées à la courbure et aux variations de courbure au sein de chaque classe. Les régions saillantes sont construites progressivement en fusionnant des classes voisines entre elles. Ce processus de fusion se poursuit tant que le degré de saillance d'une région augmente. Ce degré de saillance est défini comme une combinaison linéaire des informations disponibles dans les descripteurs des classes fusionnées, c'est-à-dire des informations liées à la surface, la courbure, le nombre de minima et de maxima locaux et la variance de courbure de chaque classe fusionnée. Dans l'approche introduite dans [Song 12b], la première étape consiste à produire une représentation multi-échelle du maillage, cf. Figure II.24. Ensuite, au lieu de simplement utiliser une somme pour combiner ces informations multi-échelles, un mo-

dèle utilisant les champs aléatoires conditionnels, *Conditional Random Fields*, CRF [7] est utilisé. La résolution de ce modèle, conformément au critère du maximum *a posteriori*, attribue une étiquette à chaque point du maillage. En règle générale, la plupart des points se voient attribuer la même étiquette, constituant les régions non saillantes tandis que les autres points forment les régions saillantes. Grâce à cette classification binaire, l'approche utilisant les CRF est plus efficace que la plupart des méthodes existantes de détection de la saillance des maillages pour capturer des saillances stables, mais la précision reste faible.

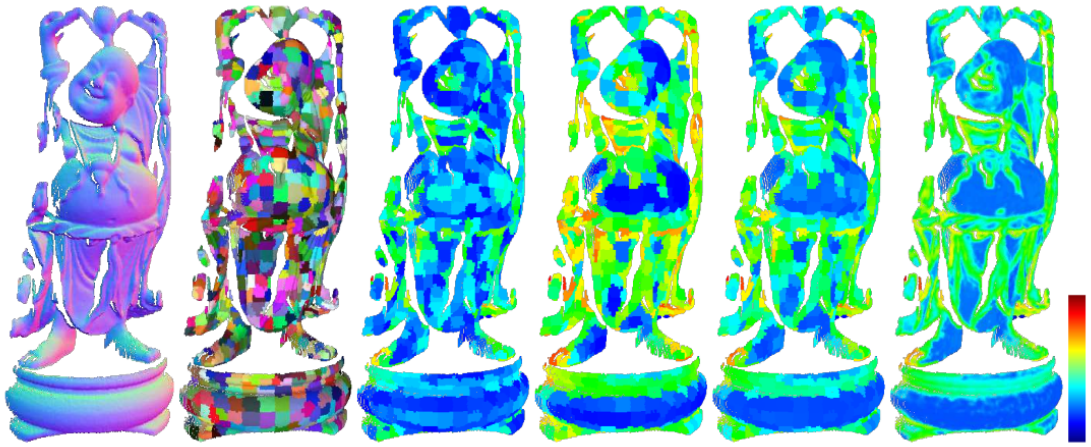


FIGURE II.25. – Étapes d'une méthode de détection de saillance 3D basée classification. L'illustration est extraite de [Tasse 15]. De gauche à droite, nous visualisons la carte des normales, la classification initiale, la classification en fonction de la distinction, la classification en fonction de la dispersion spatiale, et enfin la propagation de la saillance de chaque classe aux sommets du maillage.

De manière différente, dans la méthode de [Wu 13], les sommets sont regroupés par classes à l'issue d'une étape de segmentation. Chaque classe va être caractérisée par deux valeurs de saillance : un représentant sa distinction locale et une indiquant sa rareté globale. Plus précisément, la distinction locale d'une classe est une information locale qui indique à quel point une classe diffère de ses classes voisines, alors que la rareté globale illustre le caractère unique de classe dans la partition. La saillance de chaque sommet est ensuite calculée par une interpolation des deux valeurs de saillance de ses classes voisines, puis ces deux valeurs interpolées sont sommées. De même, dans l'approche introduite dans [Tasse 15], les sommets sont d'abord repartis en plusieurs classes à l'aide de leur version adaptative de l'algorithme SLIC, *Simple Linear Iterative Clustering* [Achanta 12], en 3D.

7. Les CRD sont des modèles probabilistes graphiques utilisés en vision par ordinateur pour modéliser et résoudre des tâches telles que la segmentation d'images, la classification d'objets, la détection d'objets, et plus encore. Un CRF modélise les relations entre des variables aléatoires observées et des variables aléatoires cachées ou non observées, en tenant compte des dépendances contextuelles entre elles.

Chaque classe se voit attribuer une valeur de saillance s'appuyant sur sa distinction locale et sa dispersion spatiale. Plus précisément, la dispersion spatiale d'une classe correspond à la moyenne pondérée des distances entre chaque point d'une classe et le centre de la classe. Les régions saillantes sont celles qui sont les plus distinctives avec une faible dispersion spatiale. Par la suite, les valeurs de saillance calculées et associées à chacune des classes sont propagées à chaque sommet qu'elles contiennent. Ces propagations s'appuient sur les probabilités d'appartenance des sommets à chaque classe. La Figure II.25 illustre les différentes étapes de cette approche. Une autre approche s'appuie sur la notion de segmentation, celle de [Tao 15], qui, à partir d'une segmentation préliminaire attribue à chaque région une valeur en fonction de sa distinction locale. Parmi les classes présentant une distinction relativement faible, certaines sont sélectionnées comme requêtes. La pertinence de chaque classe par rapport aux requêtes est ensuite utilisée pour calculer leur saillance. Enfin, la saillance de chaque classe est propagée à chaque sommet à l'aide d'un opérateur laplacien. Plus récemment, dans les travaux menés dans [Ding 19] qui sont très similaires à [Wu 13], les auteurs et autrices ont proposé un algorithme de détection de la saillance des nuages de points qui calcule séparément la distinction locale, en chaque point, et la rareté globale, de chaque classe de points. À l'aide d'une marche aléatoire, chaque point reçoit une valeur liée à la rareté globale de sa classe d'appartenance. Au lieu d'effectuer une simple somme ou une combinaison linéaire avec des coefficients fixes, une étape d'optimisation est réalisée afin de combiner au mieux ces deux valeurs pour obtenir une unique valeur de saillance finale par sommet. La fonction de coût utilisée privilégie les points ayant des caractéristiques de distinction locale et de rareté globale élevée pour obtenir une valeur de saillance plus élevée. Elle élimine également les points présentant des caractéristiques de distinction locale et de rareté globale faibles, tout en favorisant une distribution lisse de la saillance. Enfin, plus récemment, une nouvelle méthode de détection de saillance des maillages 3D colorés a été proposée dans [Ding 23]. Cette approche considère à la fois des attributs liés à la couleur et des caractéristiques géométriques. Les informations géométriques sont représentées par des descripteurs s'appuyant sur une segmentation 3D de l'objet coloré. Cette méthode a été élaborée en s'appuyant sur des expérimentations sur le suivi du regard.

4.5. Approches basées apprentissage

Étant donné que la saillance visuelle 3D concerne la perception humaine sur des données 3D, il est naturel d'envisager de l'apprendre à partir de données générées par des êtres humains. Toutes les approches basées apprentissage ont besoin de données d'apprentissage et dans ce domaine, les données d'apprentissage sont obtenues grâce à des études où on demande aux utilisateurs et utilisatrices d'indiquer les éléments saillants de l'objet. Ce qui diffère c'est la nature même de la saillance attendue.

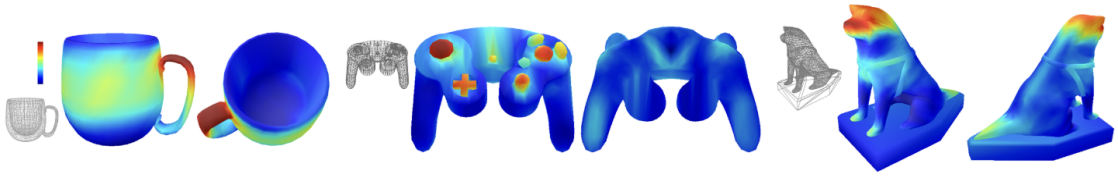
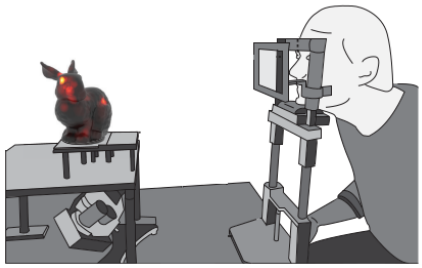


FIGURE II.26. – Illustrations du concept de saillance tactile. Cette illustration est extraite de [Lau 16]. Il s'agit de trois exemples de maillages 3D et de cartes de saillance tactiles (deux vues chacun). À gauche, les auteurs présentent un résultat de la catégorie "Saisir", une tasse alors qu'au milieu, c'est la catégorie "Appuyer", avec une manette de jeu et, enfin, à droite, la catégorie "Toucher" avec un modèle de statue. Les couleurs bleu-rouge correspondent à des valeurs de saillance relatives, le rouge étant le plus saillant.

Par exemple, dans la méthode de [Lau 16], il s'agit d'indiquer les éléments d'un objet qui sont tactiles, c'est-à-dire les zones qui sont susceptibles d'être touchées, appuyées ou saisies par les doigts, comme illustré dans la figure II.26. Ces zones ont été sélectionnées par des utilisateurs et utilisatrices lors d'une étude où il était possible de manipuler les maillages 3D des objets, par le biais d'un écran. Le modèle de régression proposé dans [Chen 12] est entraîné à partir de la distribution dite de Schelling. Plus précisément, cette distribution représente un ensemble de points sélectionnés par les participantes et participants lors d'une étude où il faut choisir les zones les plus susceptibles d'être sélectionnées par les autres personnes de l'étude. Ces données ont été, par la suite, analysées afin d'en extraire des propriétés et des caractéristiques géométriques propres aux points de Schelling obtenus. Cependant, il est coûteux et difficile d'obtenir ce genre de données, aussi, certaines techniques s'appuient sur des réseaux de neurones convolutifs faiblement supervisés comme celle de [Song 19]. Dans cette approche, on fait l'hypothèse que les objets 3D de la même classe ont généralement des distributions de saillance similaires, comme dans les publications de [Chen 12, Lavoué 18]. On peut donc étendre les annotations réalisées pour un faible nombre d'objets à l'ensemble des objets qui appartiennent à la même classe. L'originalité de la méthode de [Abid 20] réside dans le fait qu'on extrapole la saillance obtenue en 2D, ainsi qu'une correspondance entre 2D et 3D, pour extraire la saillance en 3D. Plus précisément, le modèle 3D est représenté par un ensemble d'images de rendu pour lesquelles on utilise la méthode *Salicon* [Jiang 15]. Enfin, nous pouvons citer une approche faisant appel au réseaux antagonistes génératifs ou *Generative Adversarial Network*, GAN, plutôt qu'à un réseau de neurones convolutionnels comme dans [Song 23].

Enfin, depuis une dizaine d'année, des approches d'évaluation originales ont été développées pour évaluer la qualité des cartes de saillance proposées et permettre la proposition de nouvelles approches. En effet, pour certaines approches, comme celle de [Hu 20] illustrée dans la Figure II.28, le suivi du regard ou *eye-tracking* est le principal moyen d'étudier,



(a) Schéma d'une expérience de suivi du regard

(b) Cartes de fixations du regard

FIGURE II.27. – Schéma et résultats d'une expérience de suivi du regard. Ces illustrations sont issues de [Wang 18]. En (a), on illustre l'expérience menée pour déterminer les zones où se pose l'œil humain avec la carte de fixation associée alors qu'en (b), trois exemples de cartes de fixation du regard sont présentées.

de comprendre et de valider les zones saillantes. Ces expériences, qui sont présentées dans la Figure II.27a, permettent de construire ce que l'on nomme des cartes de fixations du regard et qui représentent les zones d'attention visuelle, cf. Figure II.27b.

Nous présentons donc à présent un ensemble de techniques qui s'appuient sur l'étude du regard pour extraire des annotations de la saillance en 3D pour ensuite entraîner un modèle, comme celles de [Judd 09, Wang 18]. Nous pouvons déjà indiquer que ces études ont également permis de valider des modèles déjà entraînés comme ceux de [Jeong 17, Song 21] ou les travaux de [Wang 16] qui ont permis de vérifier si les conclusions émises dans [Yarbus 13] (c'est-à-dire que les observateurs déplacent leur regard vers les caractéristiques saillantes), s'appuyant sur les observations de stimuli en 2D, sont toujours valables pour les formes en 3D. Ils ont demandé à des observateurs et observatrices d'examiner des objets physiques (formes 3D imprimées) et ont cartographié leurs fixations sur la surface de ces formes. Ils ont constaté que, comme pour les images 2D, il existe des caractéristiques visuellement saillantes sur les formes 3D qui attirent l'attention visuelle des observateurs et observatrices. Contrairement au type de données étudiées dans [Wang 16], les auteurs de [Lavoué 18] s'intéressent aux formes 3D rendues et fournissent une analyse statistique rigoureuse de l'influence de la forme 3D et des paramètres de rendu sur les emplacements de fixation des yeux en 3D grâce à une étude utilisateurs et utilisatrices. Ils ont notamment mis en évidence l'influence de l'éclairage, du choix des matériaux et de la trajectoire de la caméra (pour le cas des scènes dynamiques) sur le choix des zones d'attention humaine.

4.6. Bilan

Les travaux présentés dans cette thèse visent à déterminer automatiquement les meilleures vues d'un objet 3D. Nous avons choisi d'utiliser des techniques traditionnelles de détection de points saillants en 3D. Cette décision repose sur plusieurs arguments.

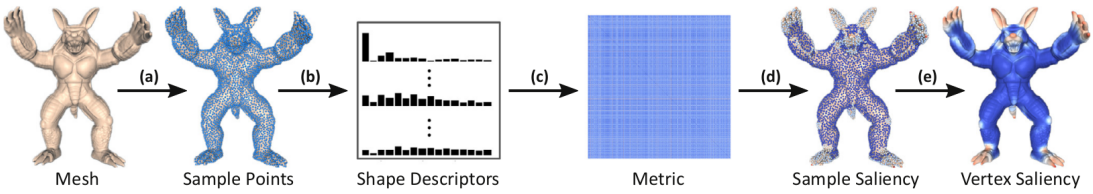


FIGURE II.28. – Approche validée par des cartes de fixation du regard. Cette illustration est extraite de [Hu 20]. En (a), l'échantillonnage des points est réalisé, puis en (b), il s'agit de la construction des descripteurs de forme pour chaque point échantillonné en s'appuyant sur la rareté globale, comme dans [Wu 13], et la faible densité. En (c), on effectue les calculs des distances entre chaque paire de descripteurs, et enfin, en (d) et (e), la résolution du problème d'optimisation de la rareté et l'interpolation de la saillance sont présentées. Il est important de souligner que les valeurs de saillance obtenues sont validées via des cartes de fixation du regard.

Tout d'abord, les techniques traditionnelles de détection de points saillants en 3D ont été largement étudiées et validées dans la littérature. Des travaux tels que ceux détaillés dans [Lee 05, Song 14, Tasse 15, Leifman 16, Limper 16] ont démontré l'efficacité de ces approches dans la mise en évidence des caractéristiques saillantes de maillages 3D. Leur utilisation permet donc de s'appuyer sur des méthodes éprouvées et bien établies. Ainsi, par la suite, nous avons choisies des techniques de saillance 3D qui couvrent une variété de méthodologies, telles que l'analyse spectrale, la classification ou encore la mesure de l'entropie.

De plus, les techniques traditionnelles offrent généralement une certaine interprétabilité et une facilité d'utilisation. En effet, contrairement aux approches basées apprentissage, les méthodes traditionnelles reposent souvent sur des principes géométriques ou statistiques interprétables, ce qui les rend plus accessibles aux personnes qui les utilisent.

Enfin, concernant notre choix de ne pas utiliser des techniques s'appuyant sur le suivi du regard dans notre méthode de sélection de meilleure vue, cela s'explique par plusieurs raisons. Tout d'abord, l'acquisition de données de suivi du regard peut être coûteuse et nécessite souvent un équipement spécialisé, ce qui rend leur mise en œuvre moins pratique dans un contexte de recherche général. De plus, les données de suivi de regard peuvent être sujettes à des biais liés aux participantes et participants et à l'environnement expérimental, ce qui peut limiter leur applicabilité dans des situations plus diverses. En outre, les données résultantes des expériences de suivi du regard ne sont pas généralisables à tous les modèles 3D ; elles sont au contraire propres aux modèles 3D étudiés, voire à leurs classes d'appartenance. En revanche, les caractéristiques de saillance intrinsèque traditionnelles sont plus robustes et peuvent se généraliser à n'importe quel maillage 3D, offrant ainsi une approche plus flexible et adaptable à une variété de situations.

En conclusion, notre choix d'utiliser des techniques traditionnelles de détection de points

saillants en 3D repose sur leur fiabilité, leur interprétabilité, leur robustesse et leur accessibilité.

En résumé

Le processus de sélection du meilleur point de vue d'un objet 3D est crucial pour comprendre et identifier correctement cet objet. Pour réaliser cette tâche, diverses méthodes ont été étudiées, chacune visant à optimiser l'expérience visuelle dans des contextes spécifiques.

Plutôt que d'opter pour des approches d'apprentissage, nous avons choisi des méthodes classiques s'appuyant sur l'extraction d'attributs géométriques du maillage 3D. Cette décision résulte de considérations telles que la disponibilité des données d'entraînement, la robustesse aux variations des données et la capacité à expliquer les critères de sélection.

L'état de l'art que nous avons présenté indique que les méthodes de sélection de points de vue utilisent divers attributs tels que la surface, la silhouette, la profondeur, la stabilité, la courbure et la sémantique.

Nous avons choisi d'utiliser et de combiner des techniques classiques de détection de saillance intrinsèque en 3D. Ce choix d'utiliser de la saillance intrinsèque, nous permet de privilégier les éléments les plus pertinents dans chaque vue, formant ainsi l'information essentielle disponible dans la vue.

Chapitre III

Points de vue restreints

Sommaire

5. Pose favorable et image 2D la plus révélatrice d'un objet 3D 63
   5.1. Contexte 63
   5.2. Intérêt de la pose d'un objet 3D 66
   5.3. Attribut révélateur d'une image 2D 70
   5.4. Filtrage de la carte de saillance multi-échelle 72
      5.4.1. Notations utilisées pour les métriques de filtrage 73
      5.4.2. Utilisation d'une distance 73
      5.4.3. Statistiques classiques 74
   5.5. Classement d'images en fonction de la mise en valeur d'un objet 3D 75
      5.5.1. Jeux de données 2D/3D 75
      5.5.2. Classement à partir des cartes de profondeur 77
      5.5.3. Classement des images révélatrices 78
      5.5.4. Limitations et perspectives 80
6. Quantification améliorée de la pertinence 85
   6.1. Problématique 85
   6.2. Méthode déterministe proposée 87
   6.3. Choix des paramètres de la méthode déterministe 88
   6.4. Score de pertinence 90
   6.5. Méthodes utilisant un score de confiance classique en apprentissage profond 92
      6.5.1. Introduction aux réseaux de neurones 92
      6.5.2. Évaluation de l'esthétique d'une image 94
      6.5.3. Lien entre apprentissage humain et réseaux de neurones . . 96
   6.6. Construction des classements de référence 98
      6.6.1. Classements intra-dégradations 100
      6.6.2. Classements inter-dégradations 102
   6.7. Protocole d'évaluation 103

6.8. Comparaisons quantitatives par type de dégradation 105
6.8.1. Dégradation *augmentation* 105
6.8.2. Dégradation *occultation* 107
6.8.3. Dégradation *changement d'échelle* 109
6.8.4. Dégradation *luminosité* 111
6.8.5. Dégradation *flou gaussien* 113
6.8.6. Bilan 114
6.9. Comparaisons quantitatives sur la combinaison de dégradations . . . 115
6.10. Résultats qualitatifs 118

Problématique

Ce chapitre, explore la problématique de l'évaluation de la pertinence d'une vue 2D présente dans une image au regard d'un modèle 3D. Cela signifie que les vues sélectionnées sont restreintes car elles sont fournies par les images disponibles. Nous avons abordé cette problématique en deux étapes. Tout d'abord en proposant de déterminer qu'elle est la pose favorable et ainsi l'image 2D la plus révélatrice ou autrement dit la plus pertinent pour visualiser un objet 3D donné. Ainsi la section 5 présente ce que nous avons mis en œuvre pour déterminer si un objet est correctement illustré dans une image spécifique. Dans un second temps, dans la section 6, nous avons considéré des caractéristiques supplémentaires dans la représentation de l'objet dans l'image et cela nous a permis d'éliminer certaines contraintes sur le choix des images testées, présentes dans la première partie.
Plus précisément, dans la première partie de nos travaux, nous avons effectué une évaluation qualitative de nos résultats, mettant l'accent sur des analyses de la pertinence des poses et dans quelle mesure une image est révélatrice d'un objet 3D. En revanche, dans la deuxième partie, nous avons réalisé une validation et des analyses plus complètes en tenant compte de nouvelles dimensions, inspirées du monde de la photographie telles que la taille et la dominance de l'objet dans l'image, et nous avons ajouté des comparaisons qualitatives et quantitatives vis-à-vis des méthodes basées apprentissage.

5. Pose favorable et image 2D la plus révélatrice d'un objet 3D

5.1. Contexte

Nous vivons à une époque où l'accès à des quantités massives de données de différentes natures, que ce soit des images, des vidéos, des modèles 3D ou des cartes de profondeur, est devenu assez simple. Des bases de données contenant des milliers d'images telles que la base *Imagenet* [Deng 09] ou *MS-COCO* [Lin 14], ainsi que des ensembles de vidéos comme la base *ToCaDa* [Malon 18], sont désormais disponibles. Certaines bases de données présentent la particularité de combiner des types de données diversifiés. Par exemple, le jeu de données *Pascal3D+* [Xiang 14] contient des images annotées ainsi que des modèles 3D correspondants aux objets présents dans ces images. Certaines bases incluent également des données de profondeur, acquises sous la forme de cartes de profondeur à l'aide de caméras de type Kinect, comme celles fournie dans [Lai 11].

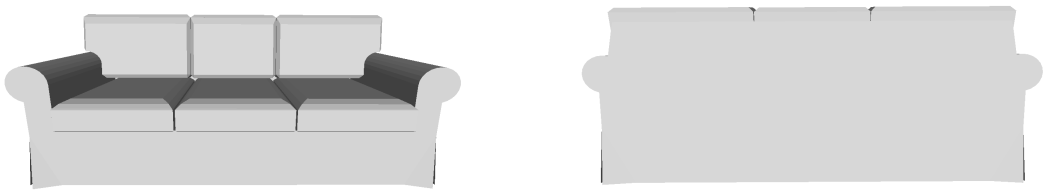
De nombreuses applications, allant de la médecine aux véhicules autonomes, nécessitent l'exploitation de données de natures différentes, pour, en général, les combiner et extraire des informations complémentaires. Par exemple, les modèles 3D, indépendants du point de vue, offrent des données géométriques 3D, tandis que les images sont liées à un point de vue

spécifique et fournissent des détails sur la texture qui sont complémentaires. D'un point de vue applicatif, nous pouvons citer, en médecine, la corrélation entre des données hétérogènes, comme par exemple entre des échographies 3D avec des images de tomographie, qui permet d'améliorer la détection précise des zones à traiter [Nam 11]. En robotique, dans le domaine de la capture du mouvement humain, certains projets [Knoop 09] intègrent à la fois des nuages de points 3D et des caractéristiques extraites d'images de caméra 2D afin de permettre un suivi robuste et en temps réel d'un être humain. De plus, la combinaison d'images 2D et de modèles 3D peut également être utile pour géolocaliser une photographie [Yang 08]. Enfin, dans le domaine des véhicules autonomes [Biglia 15], l'utilisation de données provenant de LIDAR [1] ou de RADAR [2], ainsi que de flux vidéo issus de caméras embarquées, est fréquent.

Cependant, ces vastes volumes de données, qu'elles soient dynamiques ou statiques, temporelles ou non, renferment des informations plus ou moins pertinentes. Une image, par exemple, peut contenir plusieurs objets, certains au premier plan, d'autres plus en retrait, ou moins mis en évidence. Ce qui signifie que la mise en valeur des objets peut être différente d'une image à l'autre. De plus, ces objets peuvent être partiellement visibles ou être occultés. Enfin, les objects peuvent avoir une importance plus ou moins grande pour aporter une information sur la scène et ainsi, leur contribution est limitée du point de vue de l'information globale de l'image. Nous appelons *information essentielle* la partie de l'information liée à l'image qui est indispensable à avoir pour comprendre le message transmis par cette dernière. Autrement dit, la quantité minimale d'information nécessaire pour avoir une bonne compréhension de la scène acquise. Nous faisons l'hypothèse que cette information repose sur l'ensemble des éléments saillants contenus dans l'image. Plus un objet, dans la scène proposée, sera mis en valeur, plus ses caractéristiques saillantes feront partie de l'*information essentielle* de l'image. Par définition, cette information extraite se révèle plus compacte que les données initiales, car le processus d'extraction des éléments significatifs, et le filtrage d'éléments moins pertinents, la rend plus concise.

L'objectif de nos travaux est d'identifier dans une image si un objet spécifique est bien illustré, c'est-à-dire, si nous retrouvons ses informations essentielles dans la vue offerte par l'image. Plus spécifiquement, nous nous sommes concentrées sur l'analyse d'une paire de données, composée d'un maillage 3D représentant un objet spécifique, comme représenté dans les Figures III.1a et III.1b, et d'une image 2D contenant cet objet, comme dans les Figures III.2a et III.2b. Nous disposons d'informations supplémentaires telles que la pose de la caméra et sa focale, ce qui nous permet de retrouver la projection de l'objet 3D dans l'image.

1. « Laser Imaging Detection And Ranging »
2. « RAdio Detection And Ranging »



(a) (b)

FIGURE III.1. – Illustrations de la notion de pose favorable. En (a), il s'agit d'une orientation favorable avec de nombreux détails alors qu'en (b), l'orientation est non favorable. En effet, le canapé est vu de derrière.

(a) (b)

FIGURE III.2. – Illustrations de la notion d'image révélatrice. En (a), l'image est révélatrice du canapé alors qu'en (b) elle est non révélatrice. Le canapé est globalement occulté.

Nous souhaitons être capables d'évaluer si le point de vue offert par l'image 2D est avantageux pour l'objet, en quantifiant la présence des informations essentielles de cet objet dans cette vue donnée. Le terme « avantageux » englobe deux notions distinctes :

- La pose la plus **favorable** pour un objet 3D : celle qui caractérise le mieux l'objet, dans le sens où elle permet d'obtenir une grande majorité des informations essentielles de l'objet, c'est-à-dire les éléments caractéristiques de cet objet. En effet, dans la pose de la Figure III.1a, les détails de l'assise du canapé sont visibles, alors que dans la pose de la Figure III.1b, aucun détail n'est visible, le canapé pourrait être confondu avec un autre objet. Cette notion est uniquement relative à l'objet étudié.

- L'image la plus **révélatrice** d'un objet 3D : celle qui offre une vue satisfaisante et le plus de visibilité sur les caractéristiques de l'objet étudié.L'image présentée dans la Figure III.2a est considérée comme révélatrice par rapport au canapé alors

que celle de la Figure III.2b ne l'est pas. En effet, dans l'image de la Figure III.2a, l'objet fait bien partie des éléments importants, essentiels et saillants de l'image. Il est même le principal élément avec les coussins. Alors que dans la Figure III.2b, la visibilité de l'objet est compromise en raison de la présence d'objets occultants. De plus, l'objet d'intérêt est de petite taille par rapport aux autres éléments de l'image et est positionné en arrière-plan, et les éléments saillants qui le composent ne seront donc pas visibles.

Dans la littérature, de nombreuses méthodes, comme celles détaillées dans la section 1, ont été proposées pour évaluer la qualité subjective d'une image de façon absolue, et non relative à un objet qu'elle représente. Ces méthodes s'appuient sur des critères tels que la le contraste des couleurs ou la netteté perçue [Ouni 09]. Cependant, cette évaluation est subjective et difficile à quantifier de manière objective. Nous proposons de résoudre ce problème de façon différente en considérant l'objet en 3D pour évaluer la qualité de sa représentation en 2D. Pour cela, nous nous appuyons sur des outils traditionnels en vision par ordinateur, notamment l'extraction de primitives d'intérêt.

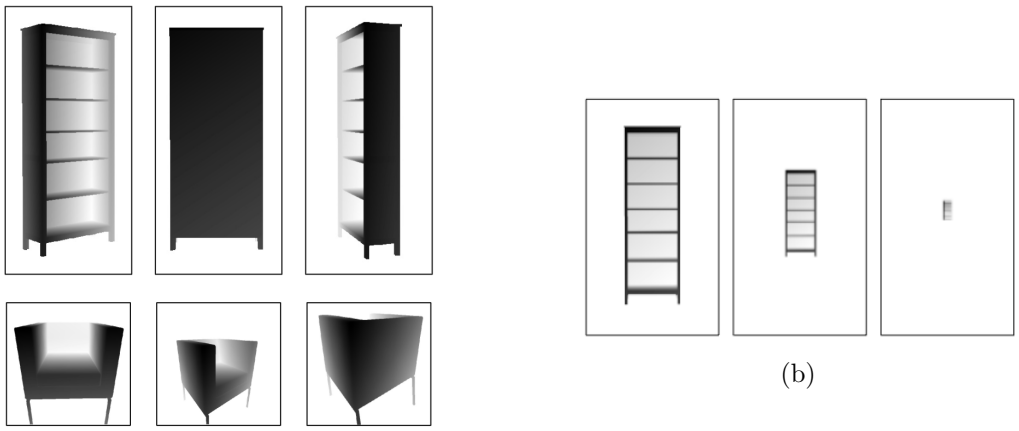
Objectif

Être capable d'évaluer la mise en valeur d'un objet 3D représenté dans une image 2D, à l'aide des éléments saillants qui constituent ce que nous appelons l'information essentielle extraite de l'image.

Dans cette section 5, nous proposons de quantifier la mise en valeur suivant deux stratégies présentées dans deux paragraphes différents. Ainsi, la première partie présente une approche utilisant les paramètres intrinsèques de l'objet vis-à-vis de sa pose au sein de l'image, caractère *favorable*, section 5.2, alors que la deuxième partie, section 5.3, porte sur l'environnement dans lequel est placé l'objet dans le sens où l'image peut être plus ou moins avantageuse pour lui, caractère *révélateur*. Les métriques utilisées pour quantifier ces deux notions sont détaillées en section 5.4. Enfin, nous présentons les résultats obtenus, en décrivant les jeux de données utilisés, cf. section 5.5.1, les expérimentations réalisées, cf. sections 5.5.2 et 5.5.3), ainsi que les limites identifiées, cf. section 5.5.4.

5.2. Intérêt de la pose d'un objet 3D

Le premier aspect considéré correspond à la notion de pose favorable et ne concerne que les modèles 3D. Pour cet aspect, nous avons choisi d'exploiter des cartes de profondeur (la pose et le calibrage sont connus). Ainsi, nous devons dans un premier temps déterminer la carte de profondeur associée à chacune des orientations étudiées pour le



(a) (b)

FIGURE III.3. – Deux critères pour évaluer la qualité de la pose d'un objet 3D. En (a), nous présentons l'influence de l'orientation. En effet, certaines parties d'un objet peuvent être cachées et en (b), l'influence de sa position qui fait qu'un objet peut apparaître plus ou moins grand.

modèle 3D. Pour cela, à partir du modèle 3D de l'objet et des paramètres de toutes les orientations que nous souhaitons générer, nous utilisons une méthode de rendu classique comme dans [Foley 94]. Le modèle de caméra utilisé est celui en trou d'épingle, modélisant une projection perspective. Les cartes de profondeur obtenues contiennent uniquement des informations géométriques relatives à l'objet étudié. Il n'y a aucun autre élément autour de lui qui pourrait l'occulter et empêcher la visualisation des caractéristiques propres à son information essentielle. contrairement aux images qui peuvent contenir des occultations. Toutefois, la quantité d'information disponible varie en fonction de la pose de l'objet dans les cartes de profondeurs : en effet, plus ou moins de détails saillants sont visibles.

Nous prenons en compte ces critères : l'orientation, illustrée dans la Figure III.3a, et la position ainsi que la position, de l'objet vis-à-vis de la caméra, représentée dans la Figure III.3b. En effet, si nous observons une bibliothèque, comme dans la Figure III.3a, nous voulons voir les étagères et non le dos de la bibliothèque, car les étagères font parties des éléments caractéristiques d'une bibliothèque. Ainsi, nous supposons que, plus il y a de points saillants visibles, à résolution constante, plus l'information essentielle est riche, et plus la pose, dans laquelle nous l'observons est avantageuse, c'est-à-dire favorable à l'objet.

En détails, comme illustré dans la Figure III.4, à partir d'un modèle 3D, nous appliquons successivement des rotations et projetons en 2D pour obtenir différentes orientations, à résolutions constantes. Pour des raisons de simplification, et parce que cela est réaliste avec le type d'objets manipulés, nous appliquons des rotations uniquement autour de l'axe vertical mais ce qui est proposé peut être généralisé à tous les axes de rotations possibles. Une fois les projections réalisées, nous utilisons la saillance curviligne CS pour extraire des discontinuités issues d'une analyse d'ordre 2 pour chacune des cartes de profondeur,

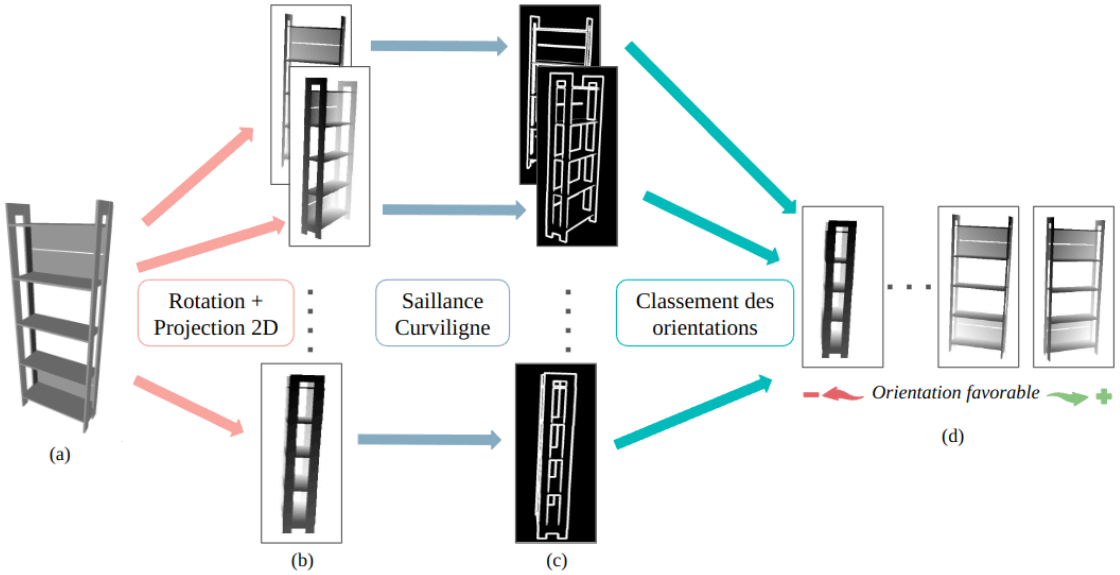


FIGURE III.4. – Chaîne de traitement pour classer les orientations favorables d'un objet 3D. À partir du modèle 3D étudié, en (a), nous générons un ensemble de cartes de profondeur de n orientations différentes, en (b). Pour chacune de ces cartes, nous calculons les cartes de saillance curviligne, en (c), qui nous permettent de proposer le classement des orientations de la plus favorable à la moins favorable, en (d).

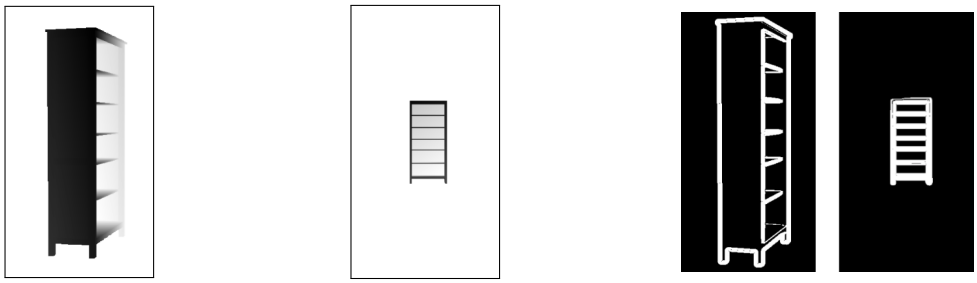
préalablement filtrées avec un filtre gaussien σ, et ainsi créer des cartes de saillance, comme illustré dans la Figure III.4(c) [3]. Plus précisément, à partir d'une carte de profondeur, nous calculons la carte de saillance curviligne CS en appliquant cette formule à tous points p :

$$CS(p) = \lambda_1(p) - \lambda_2(p) \tag{III.1}$$

avec $\lambda_2(p) \leq \lambda_1(p)$ qui représentent les courbures principales. Elles sont calculées efficacement à partir des valeurs propres de la matrice Hessienne, notée $H(p)$, de la surface représentée par la carte de profondeur, matrice diagonale d'après les calculs effectués dans [Rashwan 19]. Pour calculer la carte de saillance curviligne CS, nous lissons d'abord la surface associée en appliquant un filtre gaussien σ. Enfin, nous filtrons le bruit dû aux valeurs de CS qui sont trop proches de 0. Nous obtenons ainsi des cartes de saillance qui contiennent uniquement les éléments saillants visibles dans les cartes de profondeur. Une fois cette étape effectuée, nous classons les orientations en fonction du nombre de points saillants présents dans chaque carte de saillance. Nous rappelons que ces cartes de saillances nous permettent d'identifier les éléments caractéristiques de l'objet. Elles représentent ainsi la quantité d'information essentielle disponible.

Enfin, comme annoncé, nous prenons également en compte la position de l'objet dans

3. Les valeurs des différents paramètres sont étudiés dans la section 5.5.2.



(a) (b) (c)

FIGURE III.5. – Illustration de différentes poses. En (a), l'orientation est peu favorable alors que la position est favorable alors qu'en (b), c'est le contraire. Les cartes de saillance curviligne associées en (c) possèdent 23K (à gauche) et 13K (à droite) points saillants.

l'image. En effet, un objet, présent dans une image, peut avoir une orientation favorable, mais être éloigné, comme illustré dans la Figure III.5b. Pour réaliser l'étude de l'influence de la position d'un objet dans une image, nous utilisons la méthode décrite précédemment mais cette fois en faisant varier les positions, et non les orientations, de l'objet donné en entrée de l'algorithme. Cette chaîne de traitement est illustrée dans la Figure III.6.

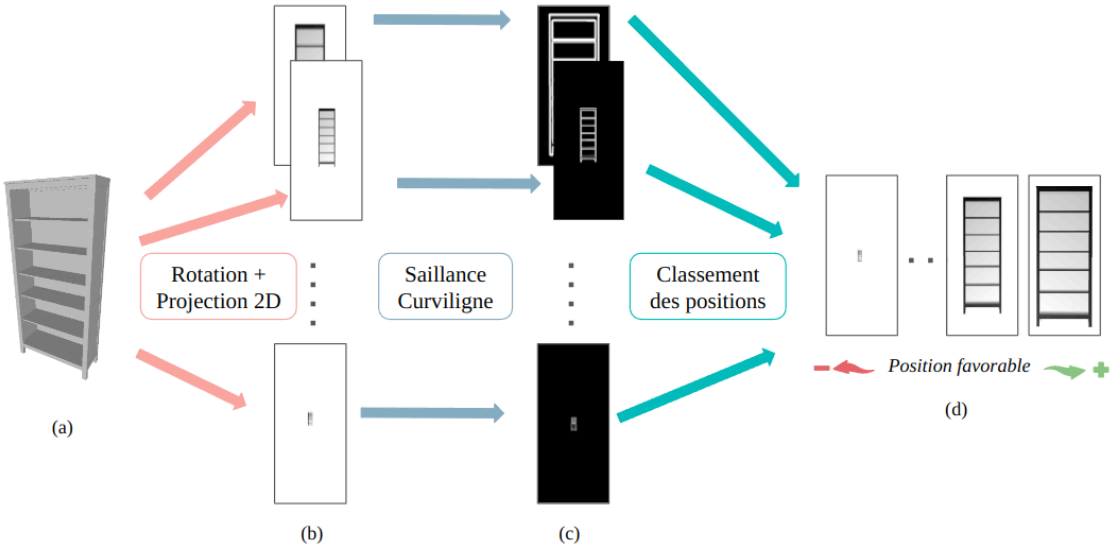
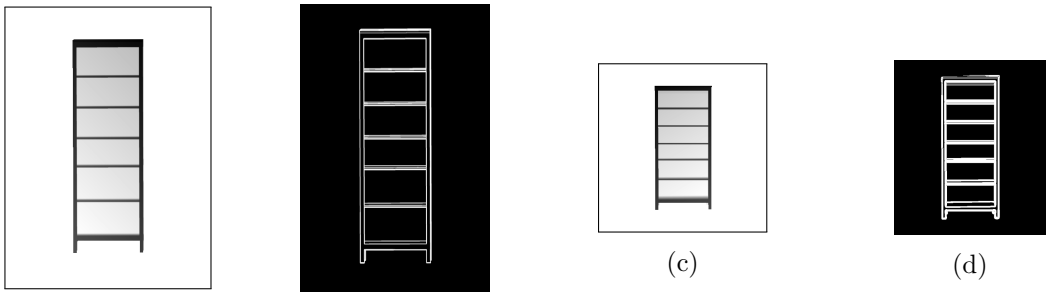


FIGURE III.6. – Chaîne de traitement pour classer les positions favorables d'un objet 3D. À partir du modèle 3D étudié, en (a), nous générons un ensemble de cartes de profondeur de n positions différentes, en (b). Pour chacune de ces cartes, nous calculons les cartes de saillance curviligne, en (c), qui nous permettent de proposer le classement des positions de la plus favorable à la moins favorable, en (d).

De même, la résolution de l'image a une influence sur le nombre de points saillants détectés et cela peut avoir un impact sur la qualité de la vue observée. En effet, pour deux

images représentant exactement la même vue d'un objet, celle de plus grande résolution contient plus de points saillants que celle de plus petite résolution. Cette observation est illustrée par un exemple dans la Figure III.7.



(a) (b) (c) (d)

FIGURE III.7. – Influence de la résolution sur le nombre de points saillants disponibles. Le nombre de points saillants détectés est plus important dans une image de grande résolution (a) que dans une image de plus petite résolution (c). Nous avons respectivement 177K et 42K points saillants dans les cartes de saillance curviligne (b) et (d).

5.3. Attribut révélateur d'une image 2D

Nous nous intéressons maintenant à l'influence des images par rapport aux objets qu'elles contiennent. Plus précisément, nous souhaitons quantifier la qualité d'une image d'un objet 3D, c'est-à-dire, à partir d'un ensemble d'images contenant le même objet 3D, identifier les images les plus révélatrices de ce dernier. Autrement dit, celles qui contiennent dans leur information essentielle le plus d'éléments caractéristiques de l'objet 3D. Intuitivement, l'image sera révélatrice si l'objet n'est ni occulté, ni tronqué et s'il appartient au premier plan. De plus, il est préférable qu'il soit placé dans une scène peu complexe pour réduire le risque que d'autres objets de la scène ne soient plus mis en avant que lui.

Le processus que nous proposons pour quantifier l'information essentielle, contenue dans une image d'un objet 3D, est illustré dans la Figure III.8. Plus précisément, pour chacune des images de l'objet, nous calculons la carte de profondeur associée et sa carte de saillance curviligne CS, comme expliqué dans la section 5.2. Contrairement à la carte de profondeur, l'image est composée de formes et de textures. Par conséquent, comme préconisé dans [Rashwan 19], nous appliquons la saillance curviligne multi-échelle MCS afin d'être robuste à la texture. Le principe de la saillance curviligne multi-échelle est d'effectuer une analyse multi-échelle qui permet de distinguer les points saillants (ceux qui ont une valeur CS élevée dans l'image) dus à la variation de la géométrie des points saillants dus à la texture. En effet, si le point saillant persiste à travers les échelles, il y a une forte probabilité qu'il ne soit pas généré par la texture mais réellement par un point caractéristique de l'image et sera conservé dans la carte de saillance finale. Par contre, si la saillance d'un

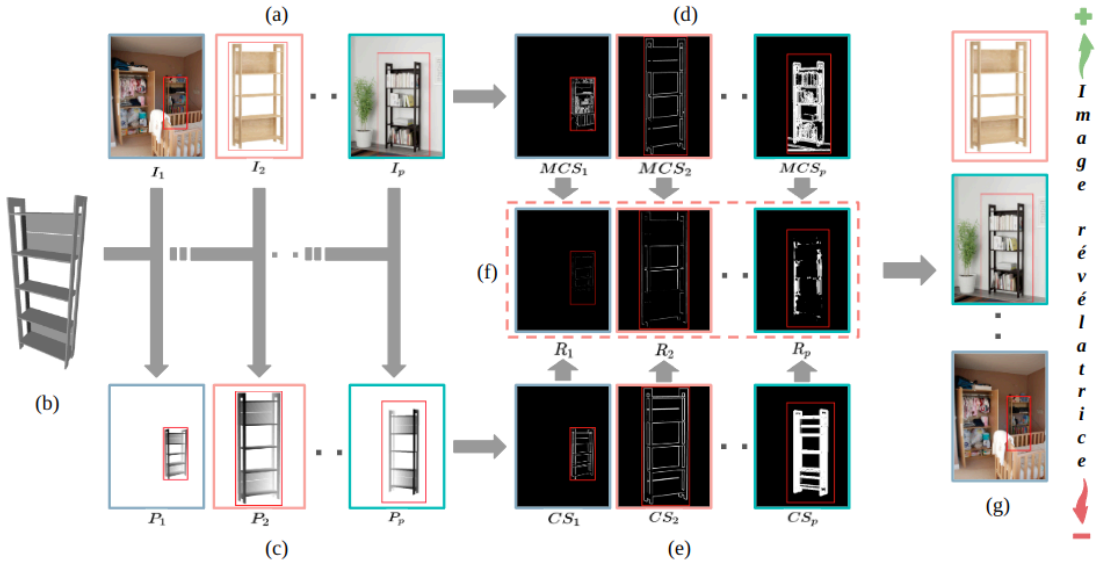


FIGURE III.8. – Chaîne de traitement pour classer les images en fonction de leur propriété révélatrice. En (a), il s'agit de l'ensemble des images notées I_i contenant le même objet et, en (b), le modèle 3D associé. Pour chaque image I_i, nous sommes capables de calculer sa carte de profondeur P_i, (c). Ensuite, nous estimons les cartes de saillance curviligne multi-échelle MCS_i associées aux images I_i, en (d), ainsi que les cartes de saillance curviligne CS_i associées aux cartes de profondeur P_i, en (e). En (f), l'extraction des points saillants appartenant à l'objet dans les cartes de réponse R_i, nous permet de proposer le classement des images de la moins révélatrice (en bas) à la plus révélatrice (en haut), en (g).

point ne s'exprime qu'à une échelle donnée, il sera considéré comme de la texture et ne sera pas conservée dans la carte finale. Les différentes échelles correspondent simplement et de manière classique à l'image filtrée par un masque gaussien combiné à une diminution successive de la résolution afin de construire une pyramide gaussienne multi-échelle et la saillance curviligne est appliquée à chaque niveau de la pyramide. Nous avons donc une carte de saillance curviligne CS provenant de la carte de profondeur de l'objet, et une carte de saillance curviligne multi-échelle MCS, provenant de l'image. À partir de ces cartes, nous obtenons un ensemble de points d'intérêt pour chaque modalité. L'étape suivante consiste à conserver uniquement les points saillants appartenant à l'objet, c'est-à-dire redondants entre l'image et la carte de profondeur. Nous supposons que la carte de profondeur nous fournit, via la carte de saillance curviligne CS, l'intégralité de l'information essentielle de l'objet alors que, dans l'image, les éléments saillants ne sont pas forcément dus à la présence de l'objet d'intérêt. Ainsi, la quantité d'information, liée à l'objet, présente dans la carte de saillance curviligne multi-échelle MCS est au plus égale à la quantité d'information disponible dans la carte de saillance curviligne CS de la carte de profondeur associée. La principale cause de points saillants non pertinents détectés dans

l'image provient de la présence d'occultations. En effet, dans cette situation, certains points saillants appartiennent à l'objet occultant mais sont exactement à la même position qu'un point saillant de l'objet étudié. Cette difficulté est illustrée dans la Figure III.9 qui affiche les cartes CS et MCS respectivement associées à une carte de profondeur et à une image. Nous observons que certains points saillants issus des plantes, par exemple, sont positionnés aux mêmes endroits que des points saillants appartenant au canapé. Nous avons donc besoin de définir un outil de filtrage pour obtenir une carte de réponse qui ne conserve que l'information essentielle de l'objet disponible dans chaque image. Nous définissons en détails cet outil dans la section 5.4. La carte de réponse ainsi filtrée est illustrée dans la Figure III.8. Enfin, à partir de ces cartes de réponse, nous extrayons les points saillants restants et proposons un classement des images en fonction de leur attribut révélateur. Plus précisément, pour chaque image, nous calculons le rapport entre le nombre de points restants dans les cartes de réponse et le nombre de points saillants disponibles dans les cartes de saillance curviligne. Si l'objet est parfaitement mis en valeur dans l'image, ce rapport vaut 1.

(a) (b) (c)

FIGURE III.9. – Carte de saillance curviligne multi-échelle, en (b), d'une image, présentée en (c) ainsi que la carte de saillance curviligne pour la carte de profondeur associée, en (a).

Comme annoncé, la section suivante indique la façon de filtrer les cartes de saillance multi-échelle afin de ne conserver que les points d'intérêt qui appartiennent réellement à l'objet d'intérêt et non au reste de la scène. Pour cela, nous détaillons les méthodes de mise en correspondance de ces deux cartes (CS et MCS).

5.4. Filtrage de la carte de saillance multi-échelle

Pour déterminer si un point saillant appartient effectivement à l'objet d'intérêt dans l'image, nous regardons son voisinage et nous le comparons à celui dans la carte de saillance curviligne CS, qui est notre référence puisqu'elle ne contient que l'information géométrique de l'objet d'intérêt. Nous faisons l'hypothèse que si nous sommes en présence de deux points homologues, alors leurs voisinages doivent être similaires, c'est-à-dire contenir une distribution de points d'intérêt équivalente. Dans notre étude, deux points sont dits correspondants lorsque les deux sont détectés à la même position, respectivement dans la carte de saillance curviligne CS et la carte de saillance curviligne multi-échelle MCS. Il

existe dans la littérature plusieurs métriques capables de mesurer le degré de similarité entre deux voisinages. Ce métriques sont variées car issues de communautés scientifiques différentes, comme la classification, la segmentation ou encore la mise en correspondance. Nous pouvons considérer différentes propriétés de robustesse de ces métriques, comme dans [Chambon 11], comme la robustesse aux bruits, aux changements de luminosité ou aux occultations. Toutefois, ici, ce qui nous intéresse c'est d'avoir une mesure qui pénalise les occultations. Ainsi, nous avons choisi des métriques de similarité classiques, simples et couramment utilisées dans la littérature au sein des communautés scientifiques citées précédemment. Nous faisons le choix de les présenter en distinguant celles qui s'appuient sur une distance, cf. § 5.4.2, de celles qui utilisent des outils de statistiques classiques, cf. § 5.4.3. Leurs performances sont analysées dans la section 5.5.

5.4.1. Notations utilisées pour les métriques de filtrage

Nous notons p_i et p_c deux points détectés respectivement dans la carte de saillance curviligne multi-échelle MCS (relative à l'image) et dans la carte de saillance curviligne CS (liée à la carte de profondeur associée). L'objectif est de déterminer si p_i appartient à l'objet dans la scène ou s'il provient d'un autre objet ou d'une texture. Nous posons $T \times T$ le voisinage carré pris en compte et celui-ci sera noté $\mathbf{V}_i$ pour le voisinage de p_i dans la carte de saillance curviligne multi-échelle MCS et $\mathbf{V}_c$ pour le voisinage de p_c dans la carte de saillance curviligne CS.

5.4.2. Utilisation d'une distance

Nous proposons d'utiliser en premier lieu une métrique très populaire en modélisation géométriques, puis dans un second temps, une métrique permettant d'être robuste aux changements d'intervalles des distributions manipulées :

- ***Distance de Hausdorff*** : permet de mesurer l'éloignement entre deux ensembles geométriques. Les deux voisinages sont préalablement binarisés : $\forall k \in \{i, c\},\ \forall\ l, c \in [1, T]^2$,

$$\mathbf{V}_k[l,c] = \begin{cases} 1 & \text{si } \mathbf{V}_k[l,c] \neq 0 \text{ (point saillant)} \\ 0 & \text{sinon.} \end{cases} \tag{III.2}$$

  Et la distance est donnée par :

$$d(\mathbf{V}_c, \mathbf{V}_i) = \max \{ \sup_{p_c \in \mathbf{V}_c} \inf_{q_i \in \mathbf{V}_i} \delta(q_i, p_c), \sup_{q_i \in \mathbf{V}_i} \inf_{p_c \in \mathbf{V}_c} \delta(q_i, p_c) \}. \tag{III.3}$$

- ***Locally Scaled Sum of Absolute Differences*, LSAD** : il s'agit d'une variante de la mesure très connue qui consiste à réaliser la somme des différences en valeur absolue, *Sum of Absolute Differences*, entre les deux voisinages $\mathbf{V}_c$ et $\mathbf{V}_i$. Cette métrique correspond à la norme L_1 localement centrées des différences. Les voisinages utilisés ne sont pas binarisés et contiennent les valeurs de saillance curviligne. Le fait d'utiliser une version centrée permet d'assurer que les deux distributions de saillance curviligne soient dans le même intervalle de valeur et donc comparables :

$$\text{LSAD}(\mathbf{V}_c, \mathbf{V}_i) = ||\mathbf{V}_c - \frac{\overline{\mathbf{V}_c}}{\overline{\mathbf{V}_i}}.\mathbf{V}_i||_1 \tag{III.4}$$

  avec $\forall k \in \{i, c\}$, $\overline{\mathbf{V}_k}$ la moyenne des intensités de $\mathbf{V}_k$.

5.4.3. Statistiques classiques

Pour les métriques suivantes, les deux voisinages sont binarisés. Voici les mesures que nous considérons car se sont les plus couramment utilisées dans la littérature :

- ***Indice de Jaccard*** : utilisé pour déterminer le degré de similarité entre deux ensembles. Il est défini par :

$$J(\mathbf{V}_c, \mathbf{V}_i) = \frac{|\mathbf{V}_c \cap \mathbf{V}_i|}{|\mathbf{V}_c \cup \mathbf{V}_i|} \tag{III.5}$$

- ***Précision-Rappel*** : cette mesure s'appuie sur les calculs de la proportions de précision et de rappel, c'est-à-dire, évaluer le pourcentage de points corrects parmi tout ceux qui ont été détectés et le pourcentage de points corrects par rapport à tous ceux qui devaient être détectés. Dans l'étude que nous proposons, nous souhaitons évaluer le nombre d'éléments pertinents trouvés en comparaison de ceux qu'il fallait trouver. C'est pourquoi, pour nous, il s'agit d'une mesure de pertinence entre deux ensembles. Ainsi, la précision est la proportion des éléments pertinents parmi l'ensemble des éléments sélectionnés tandis que le rappel est la proportion des éléments pertinents sélectionnés parmi l'ensemble des éléments pertinents.

$$\text{Précision } = \frac{VP}{VP \ + \ FP} \tag{III.6}$$

$$\text{Rappel } = \frac{VP}{VP \ + \ FN} \tag{III.7}$$

  Dans cette étude VP représente les points saillants dans les deux cartes, FP correspond aux points saillants détectés uniquement dans la carte de saillance curviligne multi-échelle et FN rassemble les points saillants détectés uniquement dans la carte de saillance curviligne. Ces derniers points sont ceux appartenant à l'objet mais qui ne sont pas saillants ou sont occultés par un autre object dans l'image.

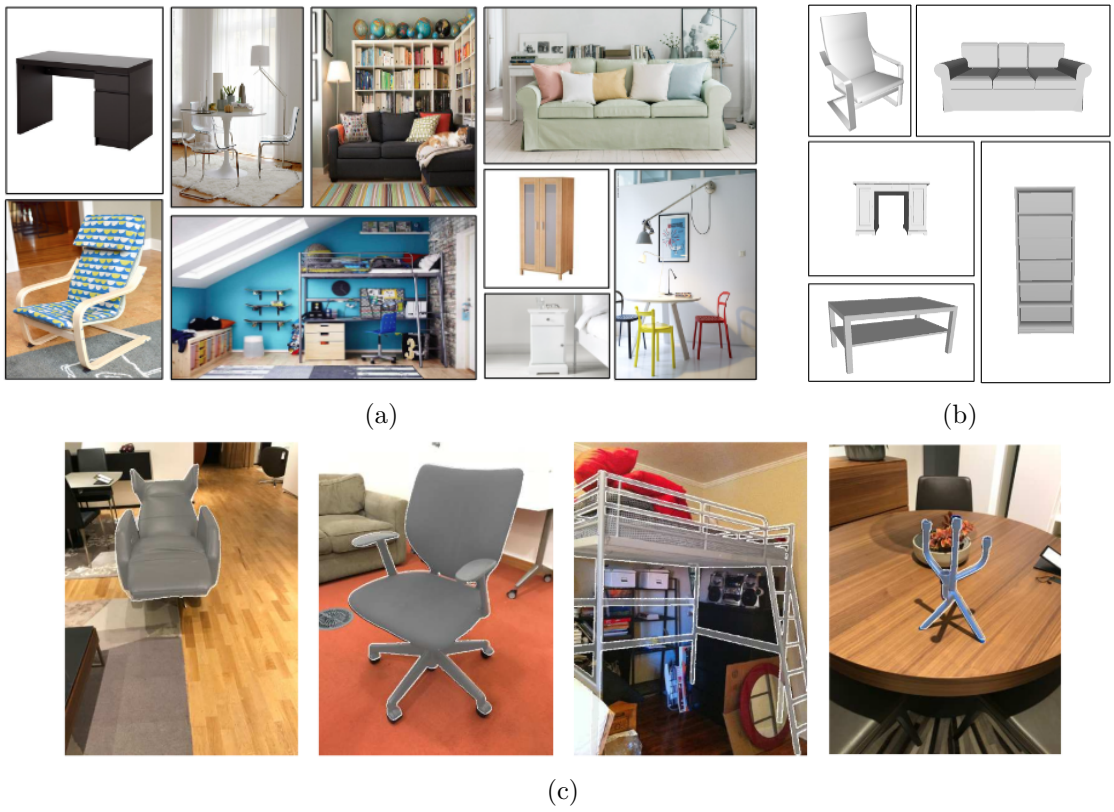
5.5. Résultats de classement d'images en fonction de la mise en valeur d'un objet 3D

Après avoir examiné les processus visant à déterminer les orientations et positions favorables pour un objet 3D (section 5.2), ainsi que le classement des images en fonction de leur caractère révélateur par rapport à cet objet (section 5.3), avec une attention particulière portée aux mesures utilisées dans le processus de filtrage (section 5.4), cette section présente la base de données utilisée. Ensuite, nous abordons la présentation et la discussion des différentes analyses effectuées, en détaillant les paramètres expérimentaux choisis, avant d'identifier les limites de notre approche.

5.5.1. Jeux de données 2D/3D

FIGURE III.10. – Mise en correspondance 2D/3D où le modèle 3D diffère de celui en 2D. Ces exemples, extraits de *Pascal3D+* [Xiang 14], illustrent les mauvaises mises en correspondance entre des images et les modèles 3D des objets qu'elles contiennent lorsque les modèles 3D ne correspondent pas exactement à ceux des objets annotés dans les images.

Pour étudier la mise en valeur d'un objet dans une image, nous avons besoin d'utiliser une base de données nous fournissant à la fois des modèles 3D et les images les contenants, ainsi que les poses des objets dans ces images. Dans la littérature, il existe un nombre important de bases de données multimodales. La base de données *Pascal3D+* [Xiang 14] offre une grande diversité avec près d'une dizaine, voire une centaine, de catégories d'objets, chacune accompagnée de dizaines de modèles 3D distincts associés à des milliers d'images 2D. Parallèlement, une autre base de données *VOC-B3DO* [Janoch 13] présente des couples d'images fournis avec leurs cartes de profondeur. Toutefois, ces deux bases de données, *Pascal3D+* et *VOC-B3DO*, possèdent des limitations : les modèles 3D sont de la catégorie des objets visibles dans les images sans être exactement le même objet, ce



(a) (b)

(c)

FIGURE III.11. – Exemples d'images (a) et de modèles 3D (b) de la base de données Pix3D. Les annotations fournies par les auteurs permettent le ré-alignement entre les objets et les images, comme illustré en (c) (illustrations extraites de l'article de [Sun 18]).

qui peut engendrer des mises en correspondance approximatives, cf. Figure III.10, et les annotations des poses peuvent être fournies avec des imprécisions, voire être incomplètes.

Pour pallier ces limitations, notre choix s'est porté sur la base de données *Pix3D* [Sun 18]. Cette dernière propose neuf catégories de modèles 3D, comprenant des objets tels que des lits, des bibliothèques, des chaises, des bureaux, des canapés, des tables, des outils, des penderies et une catégorie d'objets divers. Chaque catégorie d'objets dispose de plusieurs dizaines de modèles distincts, accompagnés de centaines, voire de milliers d'images, chacune présentant des niveaux variables de représentativité de l'objet qu'elle contient. Un exemple de données disponibles dans cette base est illustré dans la Figure III.11. Certaines images contiennent uniquement l'objet au sein d'un environnement vide, alors que d'autres montre l'objet d'intérêt dans un contexte riche, cf. III.11a. Cette base de données met à disposition un ensemble complet d'annotations, incluant les poses des caméras, ainsi que d'autres informations telles que des points d'intérêt 2D et 3D associés respectivement aux images et aux modèles 3D, les masques de l'objet d'intérêt dans les images et leurs boîtes englobantes. Les annotations ne concernent qu'un seul objet par image, contrairement à la

base de données *ObjectNet3D* [Xiang 16] qui propose des annotations de plusieurs objets par image et dont les modèles 3D associés aux images sont approximatifs. De plus, les annotations de la base de données *Pix3D* [Sun 18] fournissent un alignement précis entre les poses 2D et 3D, comme présenté dans la Figure III.11c.

La suite nous permet de présenter les résultats obtenus en considérant seulement l'objet d'intérêt, en utilisant la méthode proposée dans le § 5.2, puis en considérant une image de l'objet, en utilisant les critères proposés dans le § 5.3.

5.5.2. Classement des orientations et des positions favorables d'un modèle 3D

Dans ces expérimentations, nous avons fait le choix de générer $n = 10$ poses pour chaque objet 3D en effectuant une rotation autour d'un axe vertical. Pour le filtre gaussien utilisé pour le calcul de la saillance curviligne, nous avons choisi empiriquement $\sigma = 1,4$ qui offre un bon compromis en lissant les images sans générer un flou trop visible pour un être humain. Dans la Figure III.12, nous présentons les résultats obtenus en supposant une position et un point de vue fixe. Seule l'orientation de l'objet varie. Le classement est déterminé avec la mesure présentée dans la section 5.2. Les résultats obtenus montrent que les orientations les plus favorables sont celles avec l'objet vu de face ou éventuellement de biais. Ces observations sont conformes à nos intuitions. Si nous considérons l'exemple de la bibliothèque sur la première ligne, nous remarquons que c'est la diversité des caractéristiques visibles de l'objet qui rend son orientation avantageuse. En effet, les objets vue de dos offrent moins de détails car seuls les contours externes sont visibles.

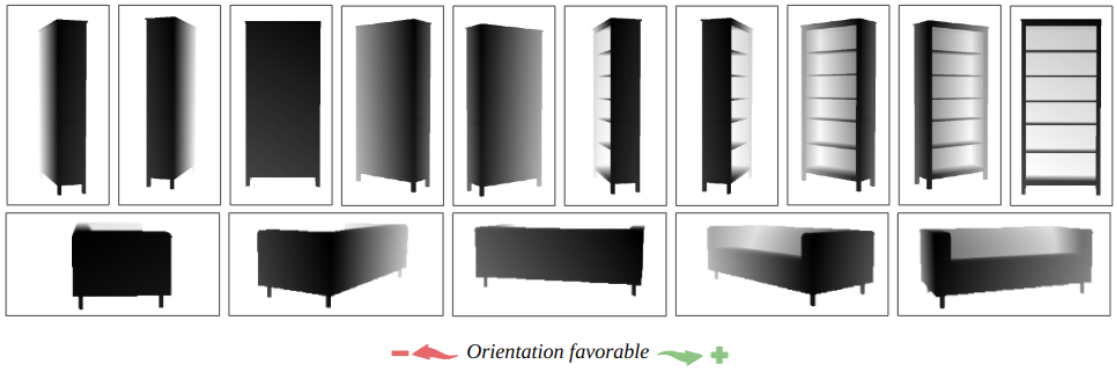


FIGURE III.12. – Classements des vues d'un même objet en fonction de son orientation. Comme attendu, les vues présentant les contours externes et internes de l'objet sont les plus favorables.

Pour la position, nous pouvons observer les résultats dans la Figure III.13. Comme espéré, ces résultats indiquent que plus l'objet est dominant dans l'image, plus il y a de points caractéristiques à détecter et plus le point de vue proposé sera favorable.

La Figure III.14 montre le classement obtenu en faisant varier l'orientation mais également la position relative de l'objet et son point de vue. Il nous semble que le classement

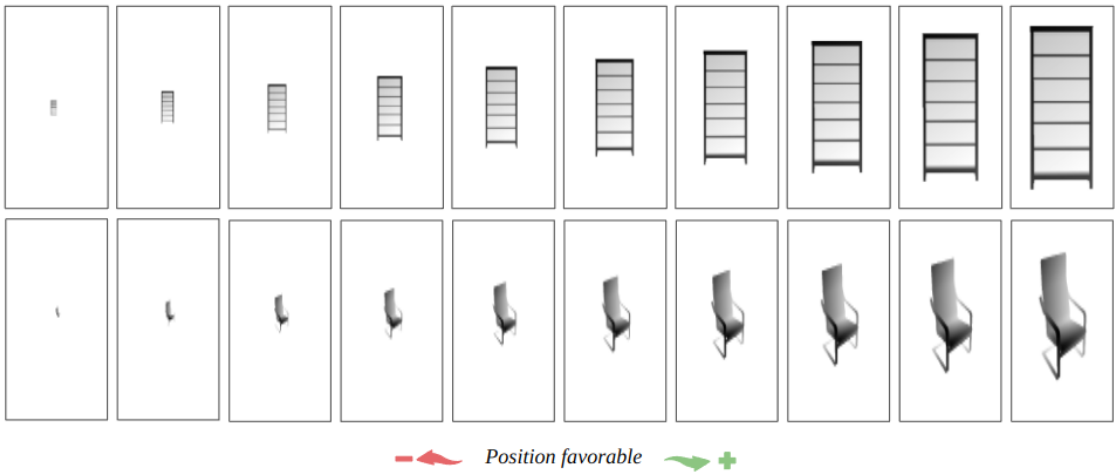


FIGURE III.13. – Classements des vues d'un même objet en fonction de sa position. De manière conforme à nos attentes, les vues les plus favorable sont celles où l'objet est le plus dominant dans l'image.

obtenu est cohérent par rapport à une qualité subjective mais nous avons besoin d'une évaluation utilisatrice et utilisateur pour confirmer ces résultats préliminaires et nous aborderons cet aspect en section 8.

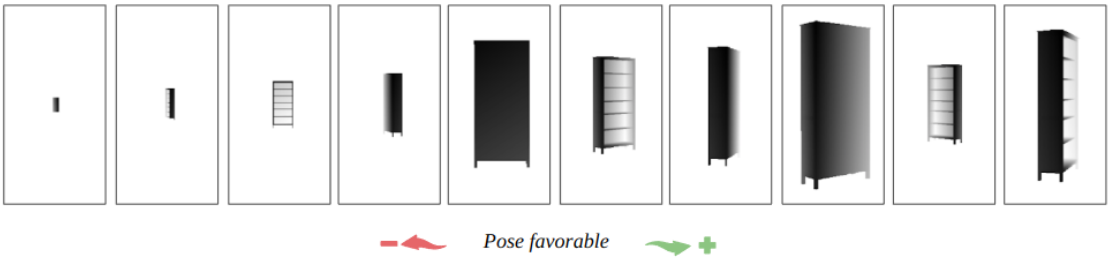


FIGURE III.14. – Classement des vues d'un même objet en fonction de sa pose. De manière conforme à nos attentes, les vues présentant les contours externes et internes de l'objet sont les plus favorables.

5.5.3. Classement des images révélatrices

Afin d'évaluer l'approche, nous avons eu besoin de construire des résultats comparables pour être en capacité de les analyser. Ainsi, nous avons choisi des ensembles d'images qui minimisent les biais induits par l'orientation, la position de l'objet et la taille des images. Autrement dit, ces trois paramètres, relatifs à l'objet, sont communs à l'ensemble des images sélectionnées pour l'évaluation. Ces choix nous permettent d'obtenir un nombre relativement homogène de points saillants dans les différentes cartes de saillance curviligne : celles issues des cartes de profondeur et celles présentes dans la pyramide d'image lors des calculs de la saillance curviligne multi-échelle. À partir des images sélectionnées

précédemment, nous avons établi des classements subjectifs en nous appuyant sur notre appréciation personnelle. L'objectif est d'évaluer, d'une part, si l'approche proposée parvient à retrouver ce résultat et, d'autre part, quelle mesure de similarité, utilisée pour le filtrage, permet de retrouver ces classements. Ainsi, par exemple, dans les Figures III.15 et III.16, nous avons convenu que l'image de droite devait naturellement être considérée comme l'image la plus révélatrice car il n'y a aucun autre objet occultant ou également saillant dans le voisinage de la table. Puis, nous avons jugé que l'image de gauche, qui présente une table avec une nappe, occultant une grande partie de la table, devait être considérée comme l'image la moins révélatrice, malgré le fait qu'elle possède une résolution plus grande que les autres images.

Pour réaliser de manière automatique le classement des images de la moins révélatrice à la plus révélatrice d'un objet, nous rappelons que nous nous appuyons sur la saillance curviligne multi-échelle présentée au § 5.3. Nous avons choisi $N_e = 4$ échelles pour le calcul de la pyramide gaussienne avec :

$$\sigma_i = \sigma k^{i-1}, \forall i \in [1, N_e] \tag{III.8}$$

où $k = 2^{\frac{1}{N_e}}$ et $\sigma = 1.4$. De plus, nous conservons, pour chaque échelle, uniquement 90% des valeurs de saillance curviligne les plus élevées. En effet, les 10% plus basses sont considérées comme du bruit. Enfin, un point saillant est conservé dans ce processus, s'il est présent au moins dans $N_c = 3$ échelles consécutives. En ce qui concerne la carte de profondeur, le traitement est le même que celui décrit dans l'étude précédente : elle est lissée avant d'appliquer la saillance curviligne.

Pour l'étape de filtrage décrite au § 5.4, nous avons choisi un voisinage de taille 9×9 et nous considérons que le point est visible si la similarité est supérieure à 70%. Ce seuil, choisi empiriquement, est celui qui permet le meilleur compromis entre maximisation du nombre de points saillants issus de l'objet et minimisation de ceux qui appartiennent à la texture ou à un autre objet de la scène. Il est également important de préciser que nous travaillons avec les boîtes englobantes et non les images d'origine. En effet, connaissant la pose de l'objet dans l'image, il est facile d'obtenir une telle boîte englobante. Cette réduction de la zone de travail nous permet juste d'être plus efficace en termes de temps de calculs. Pour pallier les imprécisions de recalage, nous avons pris la décision d'élargir cette boite englobante de 30 pixels de chaque côté.

Comme expliqué au § 5.3, en nous appuyant sur le rapport entre le nombre de points saillants présents dans la carte réponse de chaque couple (image, carte de profondeur), correspondant à l'intersection filtrée des cartes CS et MCS, et le nombre de points saillants présents dans chacune des cartes de saillance CS, nous calculons à quel point l'image est révélatrice de l'objet. Ce rapport correspond aux scores indiqués pour chacune des images dans les Figures III.15 et III.16.

Quelle que soit la mesure de similarité utilisée, la méthode proposée permet toujours de déterminer l'image la plus révélatrice suivant notre classement subjectif. De même, l'image la moins révélatrice est toujours classée en dernière position. Les trois images intermédiaires sont assez équivalentes et nous ne considérons pas leur classement comme décisif pour évaluer la qualité de l'approche. Nous souhaitions juste nous assurer qu'elles ne soient pas classées en première et dernière positions. En ce qui concerne la distance, nous avons choisi arbitrairement la distance de Hausdorff pour la suite.

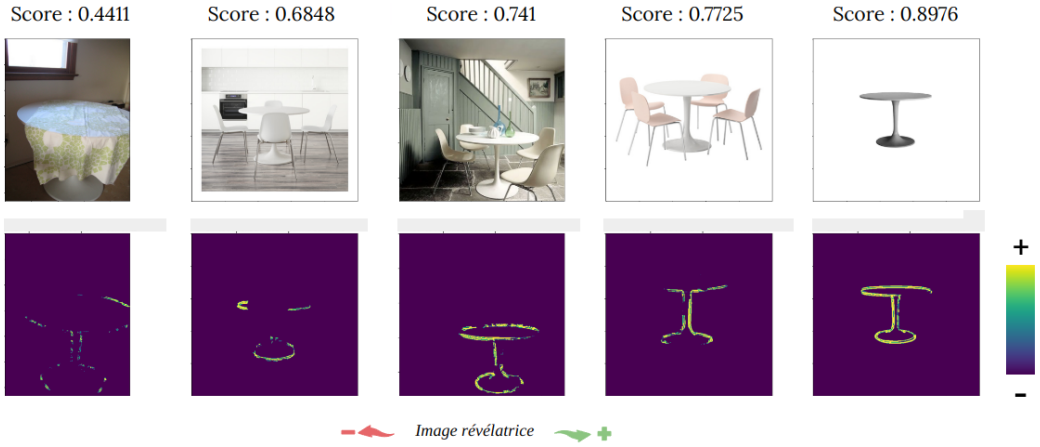
5.5.4. Limitations et perspectives

Nous avons proposé deux méthodes afin d'évaluer, d'une part, l'intérêt d'une pose d'un objet, et d'autre part, la pertinence d'une image 2D. Les résultats préliminaires obtenus sont encourageants mais nous souhaitons améliorer certains aspects.

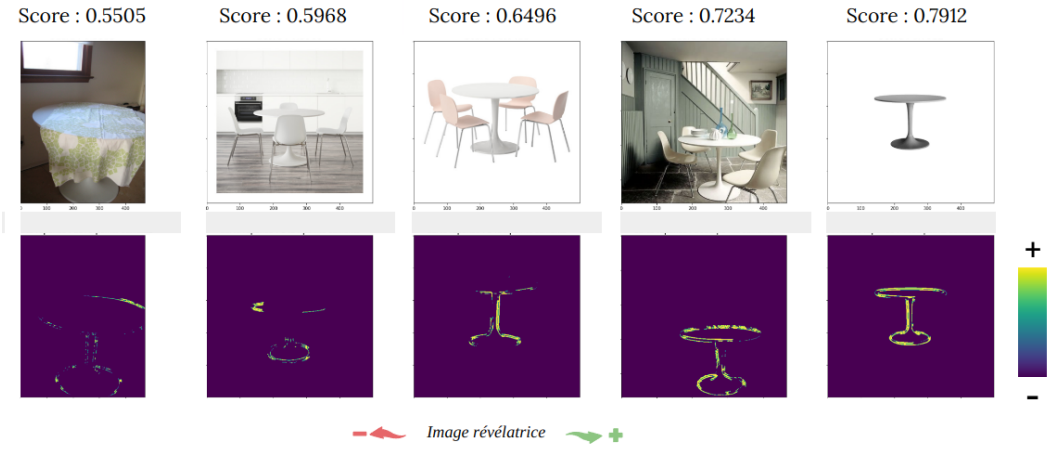
Tout d'abord, nous avons constaté que certaines cartes de profondeur ne contiennent pas toute l'information essentielle souhaitée. En effet, parfois les contours internes des objets sont de type « rampes » et présentent donc une variation d'intensité graduelle qui n'est pas détectable par les opérateurs de détection de contours et de courbure quel qu'ils soient, , comme les coins à l'intérieur du fauteuil dans la Figure III.17b. Ce manque d'information caractéristique est un biais lorsque nous déterminons le classement des poses favorables. L'ordre espéré n'est pas retrouvé. Un exemple d'un tel classement est visible dans la Figure III.17c pour le fauteuil. Avec ce manque d'informations caractéristiques d'une vue de face pour le fauteuil, par exemple, il est normal d'avoir une ambiguïté sur les orientations favorables et de ne pas retrouver l'ordre espéré. Pour remédier à cet inconvénient, nous avons amélioré notre traitement de la saillance curviligne afin de préserver une plus grande quantité d'informations provenant des cartes de profondeur.

De plus, notre approche n'est pas suffisamment robuste aux occultations. En effet, dans l'exemple de la carte de réponse de la dernière image des classements obtenus dans les Figures III.15 et III.16 il y a beaucoup de points saillants qui appartiennent à l'objet occultant (la nappe). Ces points auraient dû être rejetés.

Enfin, l'évaluation que nous avons réalisée est incomplète car nous avons supposé que l'orientation, la position de l'objet ainsi que la résolution de l'image étaient fixes. La prochaine étape est d'étudier des images en faisant varier ces trois aspects. Cette évaluation plus robuste est un des sujets de la section suivante.



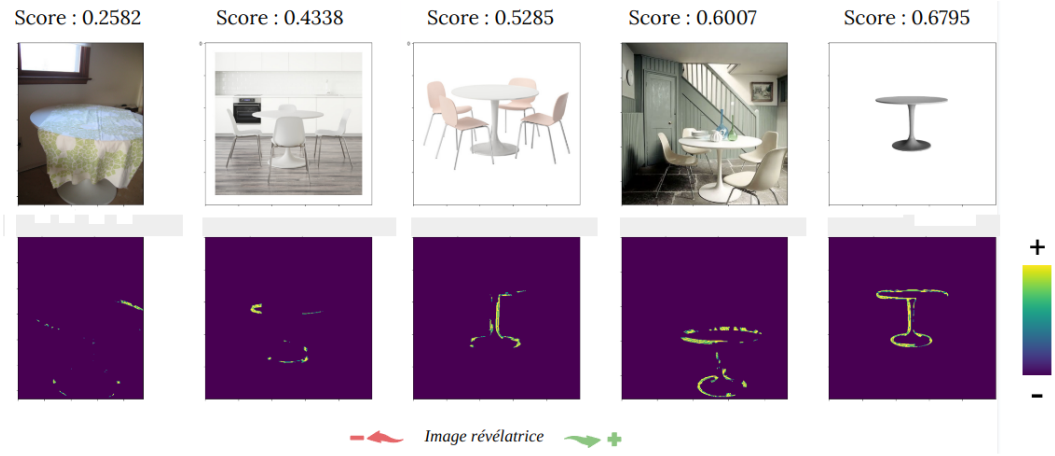
(a) Mesure de similarité d'Hausdorff



(b) Mesure de de similarité LSAD

FIGURE III.15. – Classement des images de la moins révélatrice (à gauche) à la plus révélatrice (à droite) en faisant varier la distance utilisée pour le filtrage. La première ligne présente l'image testée alors que la seconde expose la carte de réponse obtenue avec la distance d'Hausdorff en (a) et la mesure LSAD en (b). Les scores correspondent au rapport entre le nombre de points saillants présents dans les cartes de réponse, colorés en fonction de leur valeur de saillance curvilgine, et le nombre de points saillants présents dans chacune des cartes de saillance CS, indiquant à quel point l'image est révélatrice de l'objet.

(a) Mesure de similarité : indice de Jaccard



(b) Mesure de similarité : Précision-Rappel

FIGURE III.16. – Classement des images de la moins révélatrice (à gauche) à la plus révélatrice (à droite) (suivant deux outils de statistiques classiques) . La première ligne présente l'image testée alors que la seconde expose la carte de réponse obtenue avec l'indice de Jaccard en (a) et la précision-rappel en (b). Les scores correspondent au rapport entre le nombre de points saillants présents dans les cartes réponses, colorés en fonction de leur valeur de saillance curvilgine, et le nombre de points saillants présents dans chacune des cartes de saillance CS, indiquant à quel point l'image est révélatrice de l'objet.

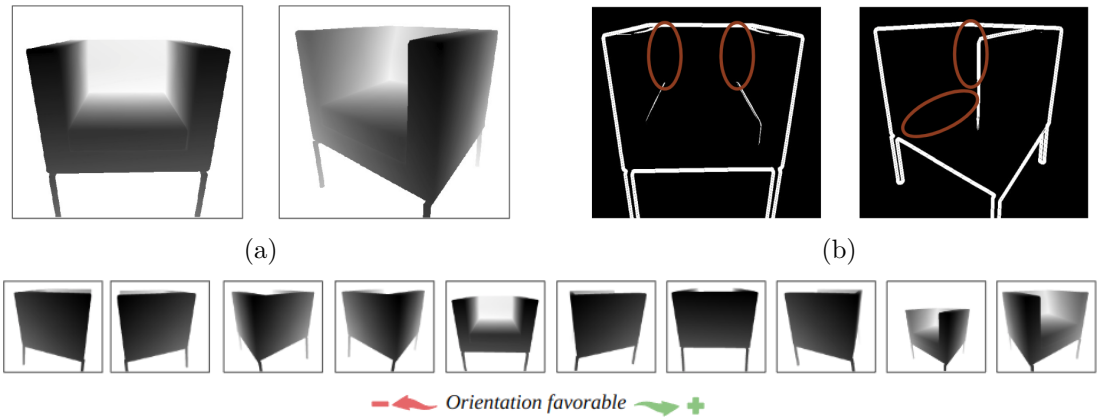


FIGURE III.17. – Illustrations des limites de l'approche déterministe proposée. Cartes de profondeur en (a), cartes de saillance curviligne en (b) avec les zones où il manque de l'information entourée en rouge.

En résumé

Dans cette section, notre objectif principal a été d'introduire une approche pour quantifier l'intérêt d'une représentation 2D d'un objet connaissant un modèle 3D de celui-ci. Autrement dit, de mesurer si l'image 2D parvient à mettre en valeur de manière satisfaisante l'objet 3D qu'elle contient. Pour cela, nous avons souhaité extraire l'information caractéristique d'un objet 3D présent dans une image 2D, en distinguant, d'une part, l'intérêt de sa pose et, d'autre part, l'intérêt de l'image 2D qui le représente. L'approche proposée s'appuie sur l'extraction de points d'intérêt dans ces deux modalités. Le détecteur choisi exploite la saillance curviligne qui permet d'assurer une répétabilité optimale entre les données 2D et 3D [Rashwan 19]. Ces travaux ont été publiés lors de la conférence nationale : ORASIS (*Journées francophones des jeunes chercheurs en vision par ordinateur*) en 2021 [Pelissier-Combescure 21].

6. Quantification améliorée de la pertinence

6.1. Problématique

Comme annoncé, notre objectif est d'approfondir et d'élargir les travaux présentés précédemment. Dans cette section, nous travaillons avec des images dont la résolution, la position ou l'orientation peuvent varier et ne sont pas prévisibles, comme illustré dans la Figure III.18. Nous continuons à nous appuyer sur l'extraction de l'information essentielle appartenant à l'objet et disponible dans l'image, mais cette fois-ci, notre approche s'enrichit en prenant en compte un ensemble plus étendu de caractéristiques liées à la représentation de l'objet dans l'image, en partie inspirées de règles classiques en photographie. En effet, nous considérons maintenant des aspects tels que la taille de l'objet représenté dans l'image et sa dominance visuelle.

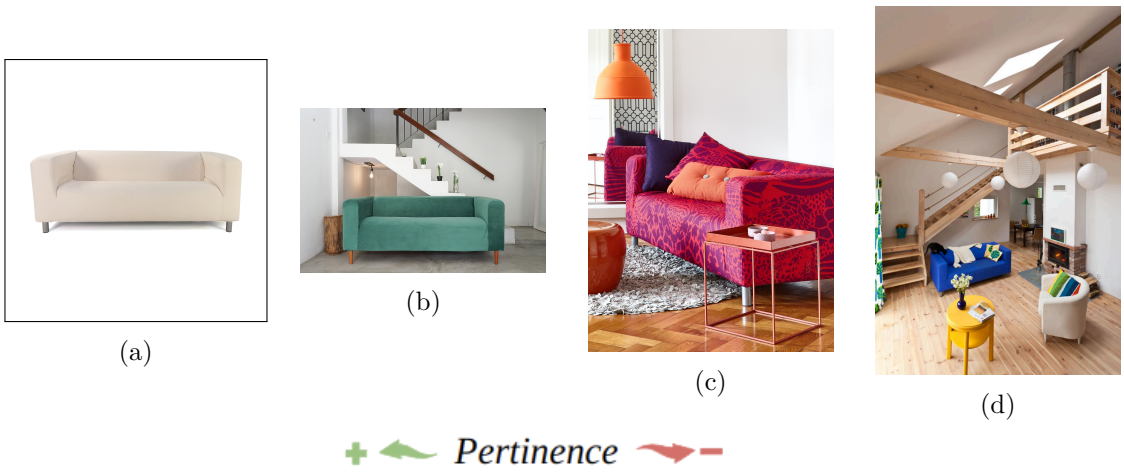


FIGURE III.18. – Classement automatique d'un ensemble d'images réelles en fonction de leur capacité à mettre en valeur un objet d'intérêt (ici un canapé) : de l'image qui met le mieux en valeur l'objet (à gauche) à la plus mauvaise (à droite).

Dans les applications médicales de réalité augmentée, par exemple, la recherche des images pertinentes pour visualiser un objet d'intérêt est indispensable, comme par exemple pour le suivi par détection où des images 2D et des modèles 3D sont utilisées dans [Collins 20]. Un autre exemple concerne les traitements. Une fois un examen réalisé, le médecin doit extraire des données pertinentes pour documenter la situation et établir le protocole de traitement approprié. Pour cela, il peut avoir besoin de constituer une collection de points de vue qui présentent au mieux la structure de l'organe et la pathologie en question. Une procédure automatique pour extraire ces vues pertinentes est d'une aide significative pour le médecin. De plus, extraire les vues les plus pertinentes parmi un grand ensemble de données, comme des images ou des images extraites d'une vidéo, est fastidieux et chrono-

phage. Il s'agit cependant d'un besoin crucial, et bien que des travaux connexes existent, aucun ne quantifie directement et automatiquement la pertinence des images d'un objet cible, indépendamment de la texture, et n'optimise ses vues.

Notre méthode permet de quantifier la pertinence d'une image sans être biaisée comme pourrait l'être un être humain. En effet, lorsqu'un humain doit choisir parmi un ensemble d'images, il introduit un biais avec sa propre perception. Il est influencé par ses préférences personnelles et ses propres goûts. Dans nos travaux, nous réalisons une quantification de la pertinence et non de l'esthétisme, autrement dit une quantification de la mise en valeur et non des préférences d'une personne.

Objectif

Quantifier la pertinence, et non l'esthétisme, d'une image 2D
pour mettre en valeur un objet 3D donné.

Ainsi, le problème consiste à identifier et à ordonner les images de la plus pertinente à la moins pertinente, comme illustré dans la Figure III.18. La dernière image met moins en valeur l'objet car il est peu dominant ou mal positionné. En revanche, les images sans environnement ou avec un objet dominant avec très peu d'occultations sont toujours, comme les images (a) et (b), placées en tête du classement.

Pour réaliser de tels classements d'images, deux possibilités ont été proposées et analysées dans nos travaux : en s'appuyant sur un « *score de pertinence* », ce qui correspond à notre contribution, ou en utilisant un *score de confiance*, inspiré des scores de confiance fournis par un réseau de neurones et qui nous permet de nous comparer à des techniques récentes. En effet, au cours des dix dernières années, les réseaux neuronaux ont continuellement démontré leur puissance et leur efficacité dans différents domaines tels que la classification ou les problèmes de régression. Ainsi, nous utilisons également un score généré par des réseaux de neurones initialement pour identifier ou classifier un objet de façon à proposer la quantification de la pertinence d'une image vis-à-vis d'un objet qu'elle contient.

Dans cette section, nous commençons par introduire la méthode déterministe proposée pour évaluer la pertinence d'une image par rapport à un objet 3D spécifique, cf. § 6.2, en précisant des détails de son implémentationc cf. § 6.3. Les paragraphes suivants se concentrent sur le calcul du score de pertinence que nous proposons, cf. § 6.4, et sur le score de confiance, cf. § 6.5, qui nous permettent d'évaluer la pertinence d'une image. Nous présentons ensuite un processus de validation, en s'appuyant sur la construction des

classements de référence, cf. § 6.6, que nous utilisons comme vérité terrain, ainsi que le protocole d'évaluation mis en place, cf. § 6.7. Enfin, nous concluons cette exploration avec une analyse des résultats, d'un point de vue quantitatif, cf. § 6.8 et § 6.9, et d'un point de vue qualitatif, cf. § 6.10.

6.2. Méthode déterministe proposée

L'intuition que nous avons reste inchangée par rapport à celle exprimée au § 5.3 : une image sera d'autant plus pertinente si l'objet étudié occupe une position centrale, n'est ni occulté ni tronqué, et apparaît clairement dans un environnement dépourvu d'autres objets. Nous voulons que la pertinence soit calculée pour un objet en fonction de sa géométrie, indépendamment de son apparence, et en tenant compte du point de vue le plus informatif. Pour déterminer la qualité du point de vue, nous proposons d'utiliser de nouveau les informations saillantes extraites dans chaque type de données, à savoir les images 2D et le modèle 3D. Cette extraction nous permet de définir la quantité d'information essentielle de l'objet disponible dans chaque image. La Figure III.19 résume le processus déterministe proposé.

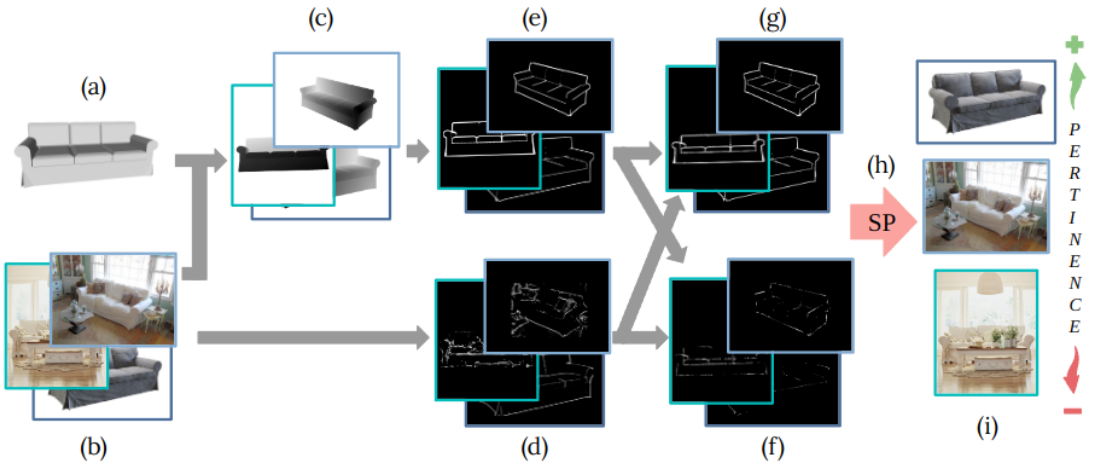


FIGURE III.19. – Chaîne de traitement pour classer des images en fonction de leur pertinence par rapport à un objet 3D. En (a), il s'agit du modèle 3D cible et en (b), de l'ensemble des images à traiter et contenant cet objet. Grâce à l'estimation de la pose, pour chaque image, nous calculons la carte de profondeur correspondante en (c). Ensuite, en (d), nous estimons les cartes de saillance curviligne multi-échelles, MCS, associées à chaque image, ainsi que les cartes de saillance curviligne, CS, associées à chaque carte de profondeur, en (e). Ensuite, nous prenons l'intersection (f) et l'union (g) entre (d) et (e). Nous ne conservons que les points saillants appartenant à l'objet. Le score de pertinence, noté SP(h), calculé pour chaque image conduit au classement de l'image en (i). Le calcul de ce score est détaillé dans le § 6.4.

Le processus global est proche de présenté au § 5.3 mais des améliorations ont été apportées, notamment au niveau de l'étape de filtrage des cartes de saillance curviligne.

Pour réaliser l'estimation de la pertinence, nous prenons en considération l'intersection les deux cartes de saillance disponibles, tout comme dans le § 5.3, mais nous ajoutons également le calcul de l'union. En effet, l'intersection entre les cartes CS et MCS pénalise les points caractéristiques de l'objet qui n'ont pas été mis en avant par le détecteur de saillance, notamment lorsqu'il y a une occultation, tandis que le calcul de l'union pénalise la localisation approximative ou erronée des points saillants. Ainsi, nous espérons que le calcul de ce rapport renforce la détection des similarités tout en pénalisant plus fortement les dissimilarités. Comme dans le processus pour le calcul d'intersection, lors du calcul de l'union, certains points saillants de l'image peuvent être dus uniquement à la texture et non à la géométrie de l'objet. Nous réalisons donc également une étape de filtrage s'appuyant sur la similarité entre deux voisinages de points homologues. Pour rappel, nous supposons que deux points sont homologues si leurs voisinages sont similaires, c'est-à-dire s'ils contiennent la même distribution de points saillants. Nous comparons donc les deux voisinages des deux points correspondants entre l'image et la carte de profondeur. Dans le cas de l'union, le filtrage s'applique lorsqu'un point est détecté uniquement dans la carte de saillance curviligne multi-échelle MCS.

À la fin de l'étape de filtrage, nous supposons que les cartes d'intersection et d'union ne contiennent que les points saillants appartenant à l'information essentielle de l'objet étudié, c'est-à-dire dus à la géométrie et non à la texture. À partir de ces deux entités, un score, que nous nommons score de pertinence, peut être déterminé pour chaque image. Nous pouvons alors classer l'ensemble des images en entrée, en fonction de ce score. Les détails relatifs aux calculs de ce score sont disponibles dans le § 6.4 mais en premier lieu, dans le paragraphe suivant, nous indiquons les choix des paramètres de la méthode proposée.

6.3. Choix des paramètres de la méthode déterministe

▷ **Carte de saillance curviligne CS :** Les paramètres pour les calculs des cartes de saillance sont identiques à ceux utilisés au § 5.5.3. L'amélioration que nous avons ajoutée ici concerne l'étape du filtrage du bruit. En effet, nous filtrons le bruit dû aux valeurs CS trop proches de 0, en éliminant tous les points de la carte CS qui ont une valeur de saillance curviligne inférieure à 1% à la valeur de saillance curviligne maximale.

$$\forall p \in CS, \begin{cases} p \text{ est conservé} & \text{si} \quad CS(p) \geq 0.01 \times CS_{max} \\ p \text{ est retiré} & \text{sinon.} \end{cases}$$

avec CS_{max} la valeur de saillance curviligne maximale dans CS.

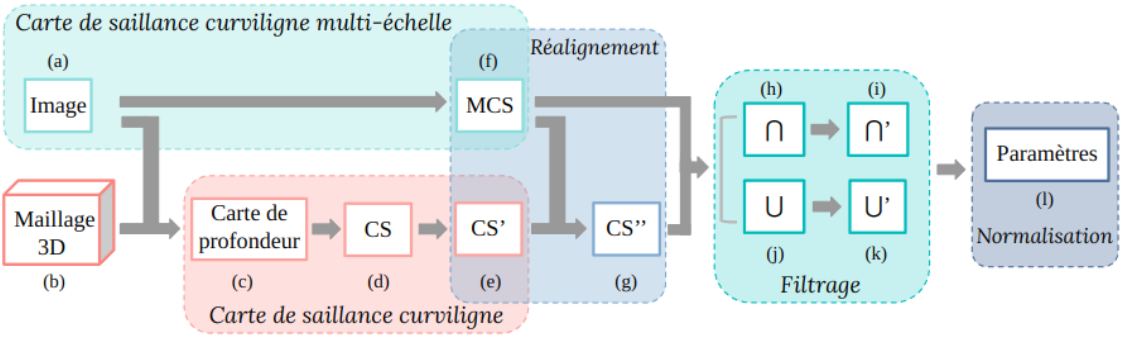


FIGURE III.20. – Processus d'estimation des scores de pertinence : En entrée, en (a), nous avons une image contenant un objet et, en (b), le modèle 3D associé. Pour pouvoir les comparer, nous calculons d'abord la carte de profondeur, en (c), à partir du modèle 3D et de la pose donnée. Ensuite, la carte de saillance curviligne multi-échelle MCS associée à l'image, en (f), et la carte de saillance curviligne CS associée à la carte de profondeur, en (d), sont estimées. Les valeurs non significatives de CS sont filtrées, en (e). Avant de comparer la carte MCS et la carte CS, un réalignement global est calculé, en (g) et, enfin, l'intersection (h) et l'union (j) sont estimées. Les versions filtrées (i) et (k) sont calculées pour générer les paramètres du score de pertinence, qui sont ensuite normalisés en (l).

▷ **Carte de saillance curviligne multi-échelle MCS :** Les paramètres utilisés sont exactement les mêmes que ceux détaillés au § 5.5.3.

▷ **Réalignement :** La base de données Pix3D [Sun 18] fournit un ensemble d'images pour chaque modèle de meubles 3D. La pose et le calibrage de chaque objet dans chaque image sont connus et nous permettent de générer les cartes de profondeur. Cependant, la correspondance entre l'image et la carte de profondeur n'est pas parfaite. Pour rectifier l'erreur de correspondance, un réalignement de la carte CS par rapport à la carte MCS est estimé afin de maximiser le nombre de points saillants présents dans l'intersection binaire. Plus précisément, les chevauchements entre la carte CS et la carte MCS translatée par le vecteur (i, j) sont calculés pour tout $i, j \in [-5, 5]$.

▷ **Filtrage :** Les paramètres utilisés dans le processus de filtrage de la carte d'union sont similaires à ceux appliqués au § 5.5.3.

▷ **Normalisation :** À la fin du processus, nous pouvons extraire les trois paramètres nécessaires au calcul du score de mise en valeur. Pour s'assurer que les trois paramètres ont la même importance, nous les avons normalisés.

Dans le paragraphe suivant, nous détaillons le calcul du score de pertinence attribué à chacune des images données en entrée.

6.4. Score de pertinence

(a) (b) (c) (d) (e) (f)

FIGURE III.21. – Différence entre un point de vue pertinent (première ligne) et un point de vue non pertinent (deuxième ligne) selon plusieurs critères. En (d), le canapé est vu de dos et tous les détails caractéristiques ne sont pas visibles, contrairement à l'orientation du canapé en (a). En (e), le canapé est occulté par de nombreux objets, et, en (f), le canapé n'est pas l'objet central de l'image.

Les photographes disposent d'un ensemble de recommandations, et non de règles fixes, pour obtenir des photographies considérées comme de bonne qualité. Ces recommandations ont été utilisées en vision artificielle pour détecter, par exemple, les zones d'attention dans les images et obtenir des cartes de saillance [Kozegar 16]. Nous nous sommes inspirées, dans une certaine mesure, de ces concepts pour élaborer un nouveau score de pertinence. Plus précisément, ce score dépend de trois termes, et nous utilisons la Figure III.21 pour illustrer nos propos :

- ψ **relatif à la notion d'information caractéristique** : Lors de la prise de vue, il est souvent recommandé de privilégier une vue qui montre le plus de détails possibles de l'objet, tout en évitant les éléments de la scène qui peuvent attirer le regard et interférer avec l'attention portée au sujet principal de la photographie. L'objectif du premier terme est d'appliquer ce concept en déterminant si l'environnement et la pose d'un objet dans une image donnée permettent d'extraire le plus d'informations caractéristiques possibles sur cet objet. Des exemples de poses plus ou moins pertinentes sont présentées en III.21a et III.21d. De plus, pour respecter cette règle photographique, l'objet d'étude doit être visible dans son intégralité, c'est-à-dire que

l'objet ne doit être ni recadré ni occulté, comme dans la l'exemple III.21e. Nous quantifions ces aspects en mesurant le nombre de points saillants visibles appartenant à l'objet étudié dans l'image : plus le nombre de points saillants est important, plus l'image est avantageuse. Plus précisément, après l'étape de filtrage, nous prenons le rapport entre le nombre de points saillants présents dans la carte d'intersection et le nombre de points saillants appartenant à la carte d'union, soit en calculant :

$$\psi = \frac{\#I}{\#U}, \tag{III.9}$$

où I, respectivement U, est l'ensemble des points saillants présents dans les deux cartes CS et MCS après filtrage, respectivement au moins une des deux.

- β **relatif à la notion de dominance** : Un objet peut montrer beaucoup de détails mais être insignifiant dans une image très grande. Deux exemples sont fournis en III.21e et III.21f. Une autre règle de photographie est de favoriser un cadrage serré afin que l'objet soit dominant dans l'image. Notre deuxième critère prend en compte cet aspect en calculant le rapport entre la taille de l'objet par rapport à la taille totale de l'image, soit plus précisément, le rapport des diagonales de sa boîte englobante de taille (w, h) et de l'image, de taille (W, H) :

$$\beta = \frac{\sqrt{w^2 + h^2}}{\sqrt{W^2 + H^2}}. \tag{III.10}$$

  Nous avons choisi une mesure linéaire car les informations caractéristiques extraites suivent des contours, c'est-à-dire des motifs linéaires, et varient donc linéairement en fonction de la taille de l'image.

- γ **relatif à la taille** : En photographie, il est également souhaité d'avoir la meilleure résolution possible. Un objet, vu dans une image, peut avoir une pose pertinente, mais être petit. De plus, pour une même dominance et orientation, la taille de l'objet, et par conséquent la résolution de l'image, a une influence sur le nombre de points saillants détectés. Plus la résolution est grande, plus il y a de points saillants disponibles. De la même façon que pour β, nous considérons une mesure linéaire pour l'objet et γ correspond à la longueur de la diagonale de sa boîte englobante, de taille (w, h) :

$$\gamma = \sqrt{w^2 + h^2}. \tag{III.11}$$

Ce terme[4] remplace la prise en compte des positions et des orientations. Il permet de maximiser la quantité d'information essentielle, et ainsi il favorise les objets dont l'image est de haute résolution.

Le score de pertinence final proposé correspond ainsi au produit de ces trois termes normalisés :

$$\mathrm{SP} = \psi \times \beta \times \gamma. \tag{III.12}$$

Nous avons choisi la multiplication des termes suite aux résultats et analyses publiées dans [Tofallis 14]. En effet, selon les observations, la combinaison par multiplication fournit une représentation plus fidèle aux préférences humaines. De plus, il est mentionné que la combinaison linéaire peut être peu fiable en raison de sa sensibilité à la technique de normalisation utilisée.

Dans le paragraphe suivant, nous présentons comment nous avons choisi d'utiliser des scores de confiance issus des réseaux de neurones pour étudier et comparer à l'approche déterministe que nous venons de présenter.

6.5. Méthodes utilisant un score de confiance classique en apprentissage profond

6.5.1. Introduction aux réseaux de neurones

Depuis 2010, grâce à une meilleure compréhension et ainsi une meilleure utilisation des fonctions d'activation, ainsi que l'accès de plus en plus aisé à de grandes bases de données et l'augmentation de la puissance de calculs, les réseaux de neurones convolutifs, *Convolutional Neural Networks*, CNN, comme LeNet, présenté dans [LeCun 98], sont devenus très populaires. En ce qui concerne le traitement d'images, c'est l'introduction du challenge *ImageNet* et l'arrivée du réseau AlexNet [Krizhevsky 12] qui ont initié l'utilisation massive de l'apprentissage profond dans ce domaine.

Des réseaux classiques permettent de résoudre un problème de classification, c'est-à-dire qu'ils produisent en sortie une étiquette de classe pour chaque image d'entrée avec un score de confiance, comme une version simplifiée de AlexNet : VGG-16 [Simonyan 14] ou le réseau Inception [Szegedy 15]. Ce dernier utilise des opérations de convolution de différentes tailles dans une architecture parallèle, démontrant que la variation des tailles des filtres peut être bénéfique pour la reconnaissance d'objets. Le réseau VGG-16 a poussé les limites de la profondeur avec des architectures comprenant jusqu'à 19 couches. En 2015, le réseau Resnet [He 16] a remporté le défi *Imagenet* après avoir introduit la notion de bloc résiduel. Le réseau DenseNet [Huang 17], en 2017, a innové en connectant chaque couche à toutes les couches précédentes, favorisant une propagation d'information plus efficace. De nombreux autres réseaux de classification ont suivi, et aujourd'hui, l'un des plus célèbres

4. Il est normalisé par la plus longue diagonale du jeu de données étudié.

est le réseau EfficientNet [Tan 19]. Ce dernier utilise une nouvelle méthode de mise à l'échelle qui ajuste uniformément toutes les dimensions de profondeur, largeur et résolution pour trouver l'équilibre idéal entre la taille et les performances du modèle. Une extension, nommée EfficientNetv2 [Tan 21], vise à améliorer l'efficacité des modèles en réduisant leur taille tout en maintenant des performances élevées et un temps d'entraînement rapide.

Depuis 2014, des architectures spécialement adaptées à la détection d'objets ont également été introduites. Au départ, il existait trois types de détecteurs d'objets : *Region based CNN*, R-CNN [Girshick 14], *Single Shot Detector*, SSDs [Liu 16] et *You Only Look Once*, YOLO [Redmon 16], ces deux derniers ayant été créés pour accroître la vitesse de calculs grâce à une stratégie de détection en une étape. À partir d'images d'entrée, ces réseaux apprennent et produisent les coordonnées des boîtes englobantes associées à des probabilités d'étiquetage de classes qui correspondent à des scores de confiance. Les premiers détecteurs ont rencontré des défis d'optimisation, conduisant à des propositions d'amélioration. Par exemple, *Fast R-CNN* [Girshick 15] a optimisé la détection en introduisant des régions d'intérêt, *Region of Interest*, *RoI*, tandis que *Faster R-CNN* [Ren 15] a automatisé la génération de ces régions en utilisant des régions d'ancrage. Une extension de ces capacités a été proposée avec le réseau *Mask R-CNN* [He 17], qui a ajouté une branche dédiée à la segmentation sémantique.

Récemment, l'introduction des transformeurs, *Transformers* [Vaswani 17], initialement proposés pour résoudre des problèmes liés au traitement du langage naturel, a marqué un tournant majeur. C'est le mécanisme d'auto-attention ou *self-attention*, SA, qui est considéré comme la clé de leur succès, permettant des interactions globales dépendantes de l'entrée, contrairement à l'opération de convolution qui restreint les interactions à une région locale. Malgré ces avantages, l'efficacité de l'auto-attention a été limitée par sa complexité, en particulier pour les entrées de haute résolution. Les transformeurs ont également gagné en popularité en vision par ordinateur depuis l'introduction du modèle *Vision Transformer*, ViT [Dosovitskiy 20]. *Notamment, une des techniques les plus avancées et qui vise à réduire la complexité du modèle initial en* utilisant des fenêtres glissantes correspond aux *Swin Transformers* [Liu 21]. Ces derniers peuvent également être exploités dans des approches semi-supervisées, comme celle présentée dans [Xu 21]. Cette dernière propose une méthode de détection d'objets semi-supervisée, basée sur un mécanisme de *soft teacher*[5]. Par la suite, des mécanismes locaux d'auto-attention ont été suggérés pour améliorer les performances et traiter les images à haute résolution, notamment avec l'introduction des *Focal Transformers* dans [Yang 21].

5. Dans le contexte des réseaux de neurones et de l'apprentissage semi-supervisé, le mécanisme du *soft teacher* fait référence à une approche où un modèle, appelé enseignant ou *soft teacher*, est utilisé pour générer des étiquettes pseudo-supervisées pour les données non étiquetées. Contrairement à l'apprentissage traditionnel où les étiquettes sont binaires, le *soft teacher* produit des distributions de probabilités sur les étiquettes pour chaque exemple non étiqueté. Ces distributions sont utilisées comme vérité terrain lors de l'entraînement du modèle principal, appelé étudiant ou *soft student*.

6.5.2. Évaluation de l'esthétique d'une image

Certains réseaux sont optimisés pour évaluer l'esthétique d'une image et une partie d'entre eux ont déjà été mentionnés dans la section 1. Ces approches récentes développent des architectures dans le but d'extraire des caractéristiques esthétiques en vue d'évaluer l'esthétique des images [Lu 14, Pan 19]. Deux catégories de méthodes se distinguent : les méthodes de classification binaire [Murray 12, Lu 14, Sheng 18] et celles de régression [Kong 16, Pan 19]. Les méthodes de régression estiment un score esthétique moyen, tandis que les méthodes de classification binaire classifient l'image comme étant de bonne ou de mauvaise qualité en établissant un seuil pour le score moyen.

La prédiction de la qualité esthétique d'une image peut être basée à la fois sur ses informations globales et locales [Lu 14]. Ces derniers ont démontré l'efficacité de leurs réseaux sur les trois problèmes suivants : classification d'une image, catégorisation de la qualité esthétique et estimation de la qualité de l'image. Ils ont élargi leur approche en présentant une nouvelle version de leur réseau dans DMAnet [Lu 15], qui, au lieu de se concentrer sur une région centrale pour extraire des caractéristiques, exploite des zones de manière aléatoire.

Une autre possibilité est d'utiliser un mécanisme d'attention pour sélectionner les zones de recherche des caractéristiques les plus saillantes [Wang 17]. Certains ont utilisé le mécanisme d'attention lors de l'étape de classification des zones de recherches [Sheng 18], attribuant un poids élevé aux zones de mauvaise qualité, tandis qu'un poids relativement faible est attribué aux zones intéressantes.

Les travaux qui estiment l'esthétique via un processus de régression intègrent à la fois des caractéristiques esthétiques, comme les ombrages et les couleurs des images, et des aspects fonctionnels tels que le flou. Les auteurs de [Pan 19] ont introduit un réseau de classification multi-tâche pour apprendre, pour une image donnée, simultanément un score esthétique et les attributs esthétiques associés, tels que la lumière, l'harmonie de couleur ou encore la symétrie. Une approche plus récente a présenté une méthode de prédiction de la distribution des scores esthétiques d'images de la même manière que le feraient des êtres humains [Xu 20]. De plus, la fonction de perte basée distance de Bhattacharyya est introduite pour calculer la similarité entre la distribution prédite par le réseau et celle fournie par une étude utilisatrice et utilisateur. Néanmoins, la plupart de ces méthodes se focalisent exclusivement sur les caractéristiques des images, sans prendre en considération d'autres sources de données pertinentes, telles que les caractéristiques textuelles provenant de légendes, comme dans [Zhang 22]. Cette approche d'évaluation multimodale de la qualité esthétique des images vise à prédire la distribution des scores esthétiques des images en entrée, avec leurs légendes respectives.

Le transfert d'apprentissage ou *transfert learning* est une technique couramment utilisée dans le domaine des réseaux de neurones. L'idée principale est d'utiliser des connaissances acquises lors de l'apprentissage d'une tâche pour améliorer les performances sur une tâche

similaire ou connexe. Cette méthode a été introduite dans [Jang 21], où les auteurs ont démontré que les caractéristiques apprises à partir de tâches de classification d'images peuvent être bénéfiques pour la classification esthétique, en effectuant ensuite un transfert d'apprentissage.

D'autres méthodes introduisent une notion d'adaptation à l'utilisateur ou l'utilisatrice, soit le concept de personnalisation. Par exemple, dans les travaux détaillés dans [Zhu 23], la personnalisation du processus d'évaluation esthétique prend en compte, pour chaque individu, ses préférences esthétiques, ses intérêts et ses expériences. L'utilisateur ou l'utilisatrice doit fournir au modèle ses préférences esthétiques sous forme de notes attribuées à un sous-ensemble d'images destiné à l'entraînement. En analysant les données relatives aux préférences personnelles, cette approche peut évaluer de manière plus précise et plus directe la perception esthétique subjective d'une image, c'est à dire, relative à une personne qui l'utilise.

Les approches basées apprentissage citées ci-dessus ont une problématique proche de la nôtre car nous étudions également la qualité de l'image. Cependant, nous nous intéressons à la qualité relative à la pertinence d'un objet 3D précis plutôt qu'à la qualité globale de l'image. Nous souhaitons attribuer un score à une image en fonction de la mise en valeur d'un objet spécifique. Une image peut donc avoir différents scores en fonction de l'objet étudié, contrairement aux réseaux de neurones qui attribuent un unique score utilisant l'esthétique globale de l'image. Ces méthodes ne sont donc pas adaptées à notre tâche. Néanmoins, l'étude que nous avons ménée sur la pertinence d'une image vis-à-vis d'un objet 3D précis peut être vue comme un problème proche de la classification ou de la régression. Il est donc légitime d'examiner le comportement d'autres réseaux de neurones par rapport à notre étude et d'observer si leurs résultats peuvent donner une réponse équivalente à la méthode déterministe proposée. Nous proposons donc une alternative au score de pertinence proposé : le score de confiance issu du domaine de l'apprentissage profond. Nous avons centré notre étude sur le comportement de deux types classiques de réseaux de neurones : les classifieurs et les détecteurs. Nous nous sommes intéressées à ces réseaux en particulier car ils ne traitent pas de l'esthétisme d'une image et peuvent avoir un lien direct avec notre objectif, que ce soit au travers d'une étiquette de classe dans le domaine de la classification ou par la détection de boites englobantes. Nous avons donc considéré que ces réseaux étaient plus adaptés, que ceux cités précédemment, pour répondre à notre problématique qui est de mesurer à quel point un objet est identifiable dans une image sans ambiguïté. Pour rappel, un classifieur, lorsqu'il reçoit une image en entrée, lui attribue une étiquette censée représenter la catégorie à laquelle appartient cette image, accompagnée d'un score de confiance qui évalue la précision du choix du réseau, comme illustré dans la Figure III.22a. Les détecteurs, quant à eux, assument un rôle similaire mais avec une étape supplémentaire en amont : la localisation des objets présents dans l'image avant de les classifier. Ainsi, pour chaque objet détecté, le réseau

fournit une boîte englobante, une étiquette correspondant à la catégorie de l'objet, ainsi qu'un score de confiance évaluant la fiabilité de cette détection, comme illustré dans la Figure III.22b. Plus ce score est élevé, plus il est probable que l'étiquette attribuée soit correcte. Notre hypothèse fondamentale de l'étude proposée est la suivante : un objet correctement mis en évidence doit être détectable et reconnaissable sans ambiguïté par un réseau de neurones.

(a) Calssifieur (b) Détecteur

FIGURE III.22. – Scores de confiance issu de l'apprentissage profond. En (a), respectivement en (b), le classifieur, respectivement le détecteur, attribue, respectivement détecte et attribue, comme étiquette à l'image, respectivement la boîte englobante, la catégorie de l'objet, avec un score de confiance p.

Nous avons choisi de considérer les modèles les plus populaires en détection d'objets et en classification. Ainsi, nous utilisons le réseau de détection d'objets YOLOv5 [Redmon 16] pré-entraîné sur la base de données *COCO* [Lin 14]. Et, pour les réseaux dédiés à la classification, nous utilisons EfficientNetv2 [Tan 21] pré-entraîné sur la base de données *Imagenet* [Deng 09].

6.5.3. Lien entre apprentissage humain et apprentissage profond par réseaux de neurones

Dans la littérature, les méthodes d'apprentissage sont devenues de plus en plus courantes dans les sciences cognitives en tant que descriptions du comportement humain, fournissant des imitations précises dans diverses tâches [Spicer 19]. L'une des questions que nous nous posons concerne la comparaison entre l'apprentissage humain et l'apprentissage automatique : à quel point ces modèles d'apprentissage peuvent être fidèles au comportement de l'être humain, c'est-à-dire à sa façon d'apprendre. Pour mesurer cette précision, des méthodes ont déjà mis en compétition des modèles basés réseaux de neurones profond, qui sont des processus mathématiques, avec les performances des êtres humains afin de déterminer si ces modèles peuvent vraiment imiter les jugements humains sur des tâches subjectives.

D'autres travaux se sont appuyés sur le concept de typicité, c'est-à-dire déterminer la caractéristique de l'objet la plus remarquée par les êtres humains, dans le but de démontrer la corrélation entre les résultats obtenus par un CNN et ceux obtenus par des utilisateurs

(a) (b)

FIGURE III.23. – Illustration du concept de typicité, extraite de [Lake 15]. Les bananes proposées en (a) sont identifiées comme étant les plus typiques selon les réseaux de neurones étudiés, tandis que celles proposées en (b) sont les plus typiques selon les êtres humains.

et utilisatrices [Lake 15]. Cependant, il est nécessaire de trouver un équilibre dans l'importance de l'imitation de la perception humaine, car celle-ci peut parfois être influencée, comme souligné dans [Barsalou 85], lorsqu'elle est altérée par une vision idéalisée de l'objet cible. Par exemple, dans [Lake 15], la représentation automatiquement identifiée comme la plus typique d'une banane par les réseaux de neurones est une banane jaune avec quelques taches brunes car c'est la plus courante au quotidien, cf. Figure III.23a. Cependant, la représentation la plus typique pour un humain est influencée par la représentation idéale qu'il en a : une banane parfaitement jaune, sans défaut, cf. Figure III.23b.

Les auteurs et autrices de [Kubilius 16] ont réalisé une expérience, permettant d'évaluer les capacités de reconnaissance par des êtres humains mais également par un processus automatique basés apprentissage profond par réseaux convolutifs. Plus précisément, ils ont utilisé un ensemble de stimuli composé d'images d'objets du quotidien avec trois variantes : images originales en couleur, images en niveaux de gris et uniquement les silhouettes. Des observateurs et observatrices ont dû nommer chaque objet, et leurs performances varient en fonction du format du stimulus. Les résultats ont également mis en évidence des similitudes et des différences dans la performance des êtres humains et des modèles pour différentes variantes de stimuli. Les CNN conservaient des performances raisonnables, même lorsque la définition d'un objet reposait uniquement sur sa forme, contrairement aux êtres humains.

Les auteurs de [Geirhos 17] ont réalisé des expériences psychophysiques afin d'évaluer la robustesse de trois réseaux de neurones bien connus, à savoir AlexNet [Krizhevsky 12], GoogLeNet [Szegedy 15] et VGG-16 [Simonyan 14], face à des dégradations d'images, en comparaison aux performances obtenus par des êtres humains, sur plusieurs tâches de reconnaissance d'objets, comme la classification pour des images présentées brièvement (200 ms), avec des variations de couleur, de contraste et bruitées. Dans une deuxième phase [Geirhos 19], une étude similaire a été réalisée de manière plus approfondie : toutes les altérations utilisées dans [Geirhos 18] ont été réutilisées puis augmentées avec plus de cinq nouveaux degrés d'intensité. L'impact de la texture sur les performances des modèles basés réseaux et des personnes humaines a également été étudié dans cette seconde phase. Les expériences menées dans [Geirhos 21], ont permis de constater que les réseaux entraînés

sur un plus grand nombre de données commettent des erreurs un peu plus proches de celles des êtres humains [Geirhos 21].

Enfin, des travaux ont été effectués pour comprendre les facteurs qui permettent d'obtenir cette similarité de comportement entre une tâche automatisée et la réalisation par un être humain. Dans [Muttenthaler 22], les auteurs ont analysé les facteurs qui affectent l'alignement entre les représentations apprises par les réseaux et les représentations mentales humaines. Ils ont constaté que l'échelle et l'architecture du modèle n'ont pratiquement aucun effet sur l'alignement avec les réponses comportementales humaines, alors que l'ensemble de données d'entraînement et la fonction objectif ont tous deux un impact beaucoup plus important.

Toutes ces observations ont confirmé la légitimité de notre décision de comparer la méthode déterministe proposée à des réseaux de neurones. Ainsi, dans le paragraphe suivant, nous élaborons une comparaison des performances des deux scores étudiés : le score de pertinence introduit précédemment, et le score de confiance, généré en sortie d'un modèle d'apprentissage profond par réseaux de neurones. Nous décrivons en détail la mise en place du protocole de validation ainsi que la construction de classements d'images considérés comme la vérité terrain.

6.6. Construction des classements de référence

(a) Image d'origine (b) *Augmentation* (c) *Occultation*

(d) *Changement d'échelle* (e) *Luminosité* (f) *Flou gaussien*

FIGURE III.24. – Illustrations des cinq dégradations considérées dans les expérimentations pour former les classements de référence permettant l'évaluation et la comparaison des méthodes.

Dans la littérature, nous pouvons trouver des bases de données qui attribuent des scores de qualité à des images. Ces scores sont établis à partir des résultats obtenus lors d'études auprès d'utilisateurs et d'utilisatrices. Toutefois, ces évaluations de qualité sont liées à des critères différents des nôtres. Par exemple, dans la base *KinIQ-10k* [Hosu 20], le score de

qualité est associé à la qualité fonctionnelle de l'image, évaluant ainsi des distorsions telles que le bruit, l'effet de crénelage ou *aliasing*, ou encore la présence d'artefacts. D'autres bases, comme *Aesthetic Visual Analysis*, *AVA* [Murray 12], fournissent un score esthétique moyen pour chaque image. Cependant, ces scores décrivent la qualité esthétique d'une image dans son ensemble, alors que nous nous concentrons sur la pertinence des images par rapport à un objet 3D en particulier. En ce qui nous concerne, aucun jeu de données n'est disponible pour évaluer les approches présentées dans cette section. Nous avons donc choisi de créer artificiellement un classement de référence que nous pouvons garantir. Pour chaque classement de référence, l'objectif des méthodes étudiées est d'évaluer la pertinence de chaque image présente dans le classement initial, fournies dans un ordre aléatoire, puis de reconstituer un nouveau classement en fonction de cette pertinence. La performance d'une méthode est plus élevée lorsque le classement généré par celle-ci se rapproche de celui de référence.

Pour créer un classement de référence, nous choisissons une image initiale à laquelle nous appliquons successivement diverses dégradations. Ces dégradations sont illustrées dans la Figure III.24. Nous avons choisi d'utiliser cinq dégradations différentes :

- *Augmentation* : ajout d'une bordure autour de l'image. La taille de l'image est modifiée contrairement à celle de l'objet. Théoriquement, seul le terme de **dominance** est affecté par cette dégradation.

- *Occultation* : ajout d'un nouvel objet par-dessus l'objet d'étude. Cette dégradation permet de tester la robustesse aux occultations garantie par l'étape de filtrage des points saillants.

- *Changement d'échelle* : diminution de la taille de l'image, ce qui induit une diminution de la résolution et donc de l'objet d'intérêt qu'elle contient. Le terme de **dominance** doit rester inchangé alors que le terme d'**information caractéristique** est très peu affecté par cette dégradation.

- *Luminosité* : modification de la luminosité qui engendre une diminution du contraste dans l'image. Nous souhaitons observer comment le terme d'**information caractéristique**, lié étroitement à la saillance curviligne, est affecté.

- *Flou gaussien* : application d'un filtre gaussien. Seul le terme d'**information caractéristique** doit être affecté, donc, l'objectif est d'étudier la robustesse de l'étape de filtrage des points saillants.

Différentes combinaisons de dégradations sont formées pour étudier le comportement et la robustesse de notre approche dans différents contextes d'applications.

6.6.1. Classements intra-dégradations

Pour étudier plus précisément le comportement et juger l'efficacité réelle des méthodes, nous avons réalisé des classements de référence par type de dégradations. Pour chacune des dégradations, nous avons formé 441 classements de référence contenant chacun 7 images : l'image initiale et six dégradations successives. Voici les détails de ces constructions :

(a) (b) (c)

FIGURE III.25. – Classement de référence contenant uniquement des *augmentations*. À partir de l'image initiale, en (a), nous avons généré l'image n°4, en (b), après 3 *augmentations* successives de 5% et l'image n°7, en (c), avec 6 *augmentations*.

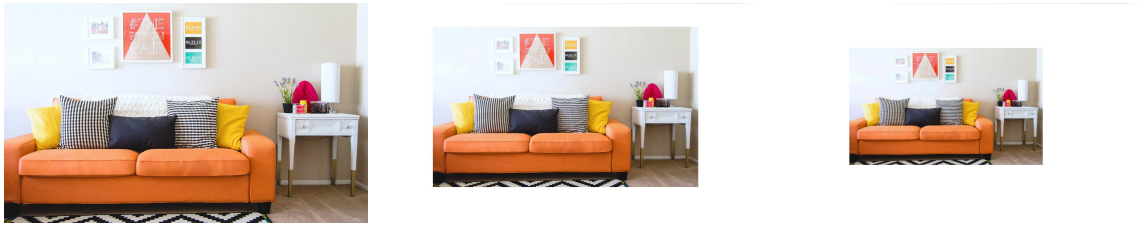
- *Augmentation* : Nous pouvons augmenter de a_w%, respectivement de a_h%, le nombre de colonnes de l'image, respectivement de lignes. Suite à l'image initiale, nous avons augmenté successivement de 5% le nombre de lignes et de colonnes sur les trois premières dégradations. Puis sur les trois dernières, nous avons modifié ce paramètre à 10%. Trois images, issues d'un de ces classements, sont affichées dans la Figure III.25.

(a) (b) (c)

FIGURE III.26. – Classement de référence contenant uniquement des dégradations de type *occultations*. À partir de l'image initiale, en (a), nous avons généré l'image n°4, en (b), après 3 ajouts et l'image n°7, en (c), après 6 ajouts.

- *Occultation* : Le paramètre *occ* nous permet d'indiquer le pourcentage à recouvrir de l'objet étudié. Dans nos classements de référence, chaque ajout recouvre précisément

10% de la boîte englobante de l'objet étudié. Ainsi, sur la dernière image, nous pouvons approximer que l'objet étudié est occulté sur 60% de sa surface visible. Trois images issues d'un de ces classements sont présentées dans la Figure III.26.



(a) (b) (c)

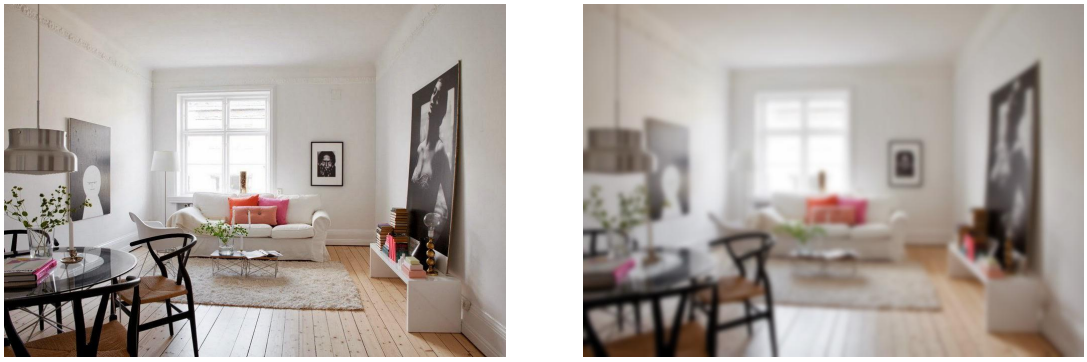
FIGURE III.27. – Classement de référence contenant uniquement des *changements d'échelle*. En (a), nous présentons l'image initiale puis, en (b), respectivement en (c), l'image dégradée n°4, respectivement n°7, ayant été réduit de 10% 3 fois, respectivement 6 fois.

- *Changement d'échelle* : à chaque itération, nous avons choisi de réduire entièrement l'image pour que le nombre de lignes, respectivement de colonnes, correspondent à 90% de ceux de l'image précédente. Ce pourcentage correspond au paramètre nommé *sc*. Trois images issues d'un de ces classements sont affichées dans la Figure III.27. Contrairement à la dégradation *augmentation*, où seule la taille de l'image est modifiée, ici, la taille de l'objet et de l'image évoluent proportionnellement.

(a) (b) (c)

FIGURE III.28. – Classement de référence contenant uniquement des changements de *luminosité*. En (a), nous présentons l'image initiale puis, en (b), respectivement en (c), l'image dégradée n°4, respectivement n°7, ayant été été altérée 3 fois, respectivement 6 fois.

- *Luminosité* : à chaque itération, la luminosité de l'image est affectée à l'aide de la fonction *RandomBrightnessContrast* de la librairie python *Albumentations* qui possède un facteur nommé l et que nous avons fixé empiriquement à 0.1. Trois images issues d'un de ces classements sont affichées dans la Figure III.28.



(a) (b)

FIGURE III.29. – Classement de référence avec uniquement des applications de *flou gaussien*. En (a), l'image initiale alors qu'en (b), il s'agit de l'image n°7 filtrée 6 fois par rapport à l'image initiale. Seules deux images sont présentées afin d'avoir une bonne visualisation.

- *Flou gaussien* : Nous avons choisi une taille de 11×11. La taille du filtre est un paramètre nommé T_f. Deux images issues d'un de ces classements sont affichées dans la Figure III.29.

6.6.2. Classements inter-dégradations

Nous souhaitons former des classements de référence contenant des combinaisons de dégradations mixtes : une version utilisant ces trois dégradations a été publiée dans [Pelissier-Combescure 23] mais une version comprenant les cinq dégradations définies précédemment, plus complète, est détaillée dans cette section. L'évaluation des performances de chaque méthode sur des classements ayant subi successivement toutes les dégradations présente un intérêt significatif. En exposant les images à une cascade de dégradations, cette approche simule des conditions réalistes où les images peuvent subir diverses altérations.

Pour construire ces classements de référence mixtes, nous avons appliqué deux fois chacune des cinq dégradations, pour obtenir 441 classements de références de onze images. Nous avons regroupé les informations nécessaires pour la création de tels classements plus complets dans la Table III.1.

Une fois que ces dégradations sont appliquées successivement, nous obtenons un ensemble d'images ordonnées de la moins dégradée (image initiale) à la plus dégradée. Au fur et à mesure que nous appliquons les dégradations de manière séquentielle, l'ordre des images est objectivement déterminé. Pour démontrer leur efficacité, les méthodes doivent retrouver l'ordre objectif de ces images.

	Dégradation	Paramètre(s)
Maillon n°1	image initiale	$\emptyset$
Maillon n°2	*Augmentation*	$(aug_w = 10\%, aug_h = 10\%)$
Maillon n°3	*Luminosité*	$lum = 0.2$
Maillon n°4	*Changement d'échelle*	$sc = 90\%$
Maillon n°5	*Occultation*	$occ = 15\%$
Maillon n°6	*Flou gaussien*	$fg = 3$
Maillon n°7	*Augmentation*	$(aug_w = 10\%, aug_h = 10\%)$
Maillon n°8	*Luminosité*	$lum = 0.2$
Maillon n°9	*Changement d'échelle*	$sc = 90\%$
Maillon n°10	*Occultation*	$occ = 15\%$
Maillon n°11	*Flou gaussien*	$fg = 3$

TABLE III.1. – Valeurs des différents paramètres lors de la création des classements de référence obtenus avec des dégradations mixtes.

6.7. Protocole d'évaluation

Le protocole d'évaluation que nous avons mis en place pour estimer l'efficacité des différentes méthodes étudiées, à partir de l'ensemble des classements de référence proposés, est illustré dans la Figure III.30.

Les images de chaque classement de référence, dont nous connaissons l'ordre, sont fournies aux trois méthodes étudiées : celle proposée qui s'appuie sur le score de pertinence, ainsi que les deux autres qui s'appuient sur des scores de confiance issus d'outils basés apprentissage profond par réseaux de neurones, YOLOv5 [Redmon 16] et EfficientNetv2 [Tan 21]. Ainsi, chaque méthode attribue un score à chacune des images. Les images peuvent alors être classées du score le plus élevé au score le plus bas. Nous possédons alors pour chaque méthode, un classement précis des images. Afin de comparer les ordres obtenus avec l'ordre attendu, nous utilisons la corrélation de Spearman, *Spearman rank order correlation coefficient*, SROCC, comme dans [Lake 15, Kong 16, Muttenthaler 22]. Plus précisément, cette corrélation utilise le coefficient de corrélation linéaire de Pearson, *Pearson Linear Correlation Coefficient*, PLCC, avec des variables de rang. Il est important de souligner que nous calculons la corrélation entre les ordres des images et non pas entre les valeurs des scores obtenus par chaque image. La mesure SROCC est définie par :

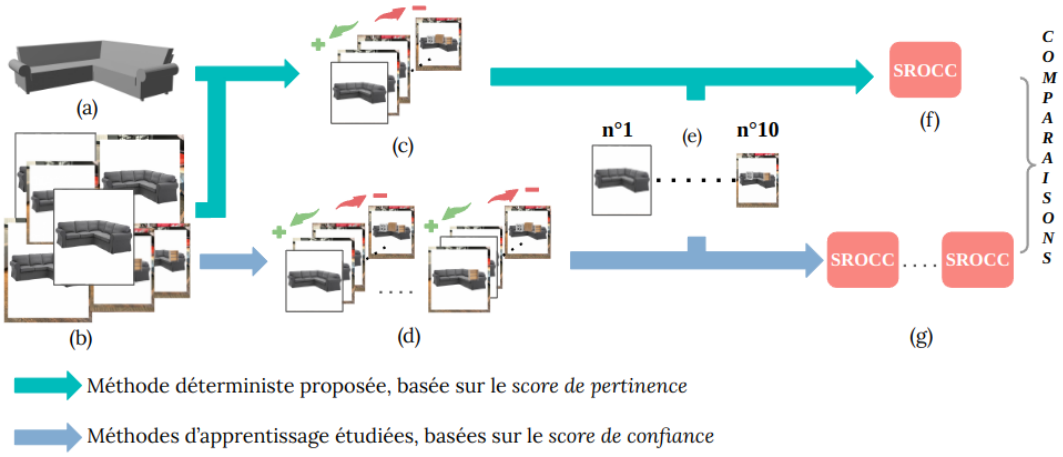


FIGURE III.30. – Évaluation des méthodes de classement. En (a), il s'agit du maillage 3D de l'objet, et, en (b), les images contenant cet objet. En (c) et (d), les classements établis respectivement avec le score de pertinence et les scores de confiance (du plus haut au plus bas) sont présentés. Enfin en (f) et (g), le coefficient de corrélation de Spearman est calculé entre le classement de référence en (e), et chaque classement obtenu en (c) et (d) afin de déterminer la méthode la plus efficace.

$$SROCC(X,Y) = \frac{cov(R(X),R(Y))}{\rho_{R(X)} \cdot \rho_{R(Y)}} \tag{III.13}$$

avec :

- $R(X)$ variable de rang,
- $cov(R(X),R(Y))$ covariance de $R(X)$ et $R(Y)$,
- $\rho_{R(X)}$ et $\rho_{R(Y)}$ écart-type de $R(X)$ et $R(Y)$.

Les valeurs de SROCC varient entre -1 et 1. Plus ces coefficients de corrélation sont proches de 1, plus la méthode est efficace. En effet, cela signifie que les deux classements, celui attendu et celui obtenu, sont similaires.

Par ailleurs, chaque coefficient de corrélation est accompagné d'une valeur appelée *p-value* : plus cette valeur est petite, plus le coefficient de corrélation est significatif. Dans la littérature, la valeur *p-value* est comparée à un seuil appelé α, généralement égal à 5.10^{-2}, ou 1.10^{-2}. En conséquence, nous ne conservons que les classements présentant une corrélation significative.

Afin de comparer les trois méthodes retenues, pour un classement de référence donné, si l'une des approches testées fournit un classement avec un coefficient de corrélation non significatif, alors cet ensemble d'images et son classement de référence associé ne seront pas pris en compte pour la comparaison entre les approches. Par conséquent, plus le nombre de méthodes à comparer dans une expérience est élevé, plus il y a de chance d'obtenir

des coefficients de corrélation non significatifs et donc d'y avoir moins de classements de référence pris en compte.

6.8. Comparaisons quantitatives par type de dégradation

Dans chacune des études de performance, détaillées ci-dessus, nous avons utilisé les mêmes conditions expérimentales. Plus précisément, pour chaque catégorie de classements de référence, nous avons étudié le comportement des trois méthodes de manière indépendante, deux par deux, puis simultanément, avec deux valeurs possibles pour le seuil α associé au paramètre *p-value* : $\alpha = 5.10^{-2}$ et $\alpha = 1.10^{-2}$. Comme expliqué précédemment, ces conditions ont un impact direct sur le nombre de classements significatifs retenus, et ces nombres sont donc précisés dans tous les tableaux de résultats présentés par la suite.

Les synthèses statistiques des coefficients de corrélation, pour les classements obtenus en utilisant le score de pertinence ou le score de confiance, sont présentées sous la forme de diagrammes en boîte. Sur ces diagrammes, les astérisques représentent le niveau des différences significatives entre la moyenne déterministe et la moyenne de chaque méthode d'apprentissage. Ces informations proviennent d'un test HSD, *Honestly Significant Difference*, de Tukey [Nanda 21]. Nous avons appliqué ce test sur nos ensembles de coefficients bruts pour vérifier si les moyennes obtenues sur chaque ensemble étaient bien significativement différentes l'une de l'autre. Ce choix de représentation nous permet de formuler facilement des conclusions sur l'efficacité relative de chaque méthode étudiée. En effet, la médiane, en tant que mesure de tendance centrale, indique la valeur au milieu de chaque ensemble de données et offre une mesure robuste de la moyenne des résultats obtenus. Une médiane plus élevée suggère généralement une meilleure performance de classification. Couplée à l'information sur la moyenne, nous avons une vue globale car nous pouvons identifier la présence ou non de résultats extrêmes. Les quartiles $Q1$ et $Q3$ viennent alors compléter cette information car ils permettent de comprendre la dispersion des données. En effet, un écart interquartile, *Inter-Quartile Range*, IQR, faible indique une cohérence dans la performance, tandis qu'un IQR important révèle une variabilité de la qualité des résultats importante. Enfin, les valeurs minimales et maximales donne ma mesure des scénarios extrêmes où chaque méthode est performante ou, au contraire, rencontre des difficultés pour fournir un résultat de qualité.

6.8.1. Analyses des performances par rapport à la dégradation *augmentation*

Nous pouvons observer, grâce à la Figure III.31a, que la méthode déterministe proposée présente des performances globalement élevées, avec une moyenne de 0.98 et une médiane de 1, indiquant une cohérence dans les résultats. Le réseau EfficientNetv2 montre une variabilité plus significative, avec une moyenne de 0.23 et une médiane de 0.82.Le réseau YOLOv5 a des valeurs inférieures pour toutes les mesures de tendance centrale, avec une

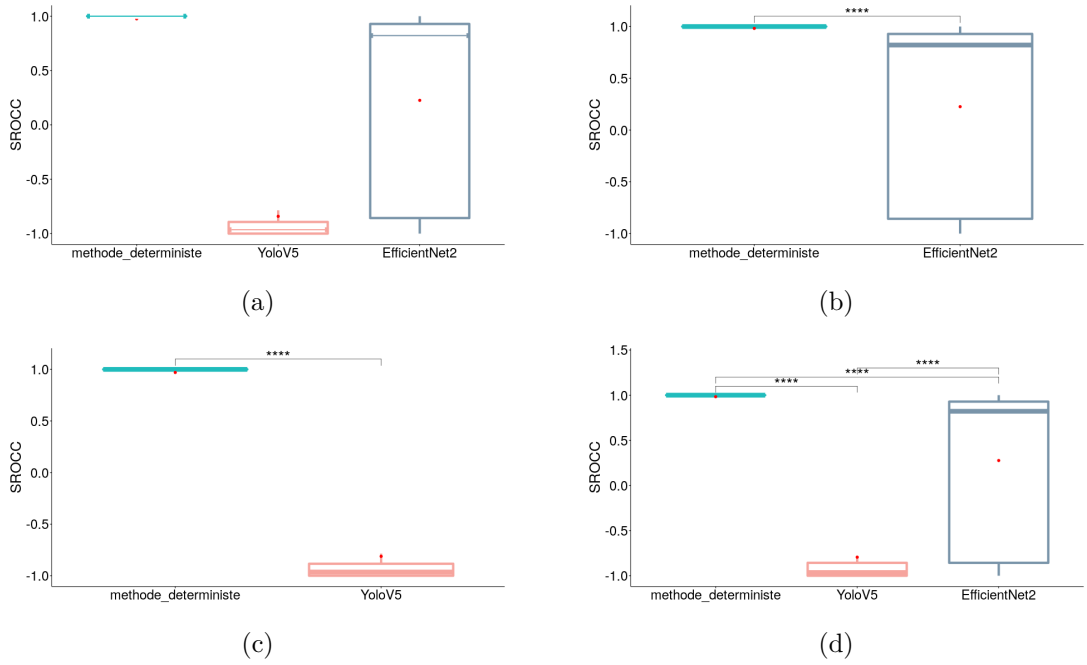


FIGURE III.31. – Résultats associés à la dégradation *augmentation*. Nous présentons les performances de chaque méthode en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expérimentations sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

moyenne de -0.84 et une médiane de -0.96, indiquant des performances généralement plus faibles.

Ensuite, lorsque nous comparons la méthode déterministe proposée directement avec chacun des deux réseaux de neurones, sur le même ensemble de classements significatifs, les observations restent similaires à celles formulées ci-dessus. En effet, d'après les diagrammes de la Figure III.31b, le premier quartile $Q1$ de la méthode déterministe proposée, valant 1, dépasse celui de la EfficientNetv2 (qui vaut -0.86), indiquant que 75% de nos classements déterminés sont identiques aux classements de référence attendus. En ce qui concerne la comparaison avec YOLOv5, avec 136 classements significatifs, les observations sont similaires. De nouveau, d'après la Figure III.31c, 75% des classements obtenus avec la méthode déterministe proposée sont identiques à ceux attendus alors que 75% de classements obtenus avec YOLOv5 ont un coefficient de corrélation inférieur à -0.88. Les classements ainsi obtenus par YOLOv5 sont totalement différents de ceux que nous avons construits pour notre validation.

Enfin, nous procédons à une comparaison globale entre les trois méthodes dont les résultats statistiques sont présentés dans la Figure III.31d. Nous pouvons remarquer que la méthode déterministe proposée montre toujours des performances de qualité avec une

$\alpha = 0.05$

	Indépendante	2 par 2		Globale
EfficientNetv2	148	108		49
Méthode proposée	309		136	
YOLOv5	209			

$\alpha = 0.01$

	Indépendante	2 par 2		Globale
EfficientNetv2	77	56		18
Méthode proposée	294		98	
YOLOv5	160			

TABLE III.2. – Coefficients de corrélation associés à la dégradation *augmentation*. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaisons effectuées parmi les 441 classements de références disponibles.

moyenne très élevée et une médiane de 1. Cela indique une cohérence dans la proximité de ses classements obtenus avec les classements de référence. La méthode YOLOv5 maintient une moyenne et une médiane plus basses, -0.79 et -0.96, respectivement. L'amplitude des valeurs reste similaire à la comparaison précédente. La méthode EfficientNetv2 se positionne entre les deux, avec une moyenne de 0.28 et une médiane de 0.82.

Les tendances demeurent cohérentes lorsque nous rendons notre sélection de classements significatifs plus stricte, en fixant la valeur du seuil α pour la *p-value* à $\alpha = 1.10^{-2}$. D'après la Table III.2, lors de l'étude indépendante, la méthode déterministe proposée conserve quasiment le même nombre de classements significatifs, nous passons de 309 à 294 coefficients de corrélation significatifs. Cependant, les approches s'appuyant sur un score de confiance sont plus affectés par cette restriction : EfficientNetv2, respectivement YOLOv5, conserve seulement 52%, respectivement 76%, de ces classements significatifs.

En conclusion, face à la dégradation *augmentation*, la méthode déterministe proposée est la plus performante.

6.8.2. Analyses des performances par rapport à la dégradation *occultation*

Pour cette étude, nous évaluons les performances des trois méthodes étudiées face à la dégradation par *occultation*. De manière indépendante, le réseau EfficientNetv2 et la méthode déterministe proposée présentent des performance quasi Notre méthode affiche une moyenne de 0.84, une médiane de 0.96, et un IQR de 0.07 et EfficientNetv2 présente des valeurs similaires avec une moyenne de 0.82, une médiane de 0.96, et un IQR de 0.11. Le réseau YOLOv5, avec une moyenne moins élevée de 0.16 et un IQR de 1.71, semble

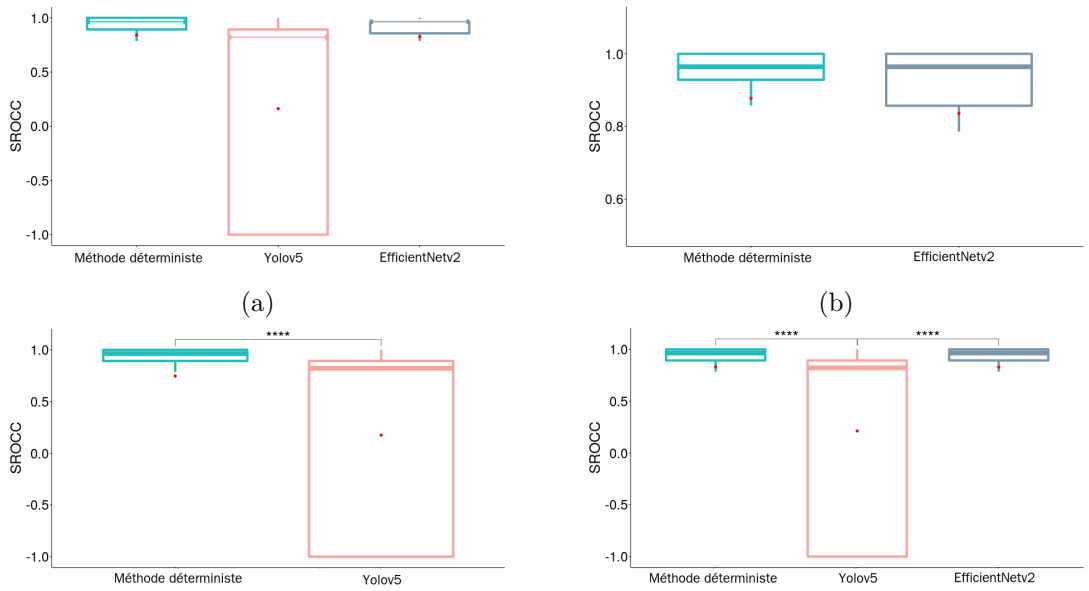


FIGURE III.32. – Résultats associés à la dégradation *occultation*. Nous présentons les performances de chaque méthode en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expériences sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

plus affecté par les occultations que les autres approches, comme l'illustre le diagramme de la Figure III.32a. Cette similitude de comportement entre le réseau EfficientNetv2 et la méthode déterministe proposée se conserve au travers des différentes comparaisons, comme le démontrent les diagrammes des Figures III.32a, III.32b et III.32d.

Les tendances restent constantes lorsque nous renforçons la significativité de notre sélection de classements, en ajustant la valeur seuil α pour la *p-value* à $\alpha = 1.10^{-2}$. Selon les résultats présentés dans la Table III.3, dans le cadre de l'étude indépendante, la méthode déterministe proposée maintient presque le même nombre de classements significatifs, passant de 323 à 282 coefficients de corrélations significatifs. En revanche, les réseaux de neurones sont davantage impactés par cette restriction : EfficientNetv2, respectivement YOLOv5, voient le nombre de leurs classements significatifs diminuer de 24%, respectivement 38%. Cette différence d'impact est la seule distinction en termes d'efficacité entre la méthode déterministe proposée et le réseau EfficientNetv2.

La méthode proposée et exploitant le score de pertinence a produit d'excellents résultats, démontrant une grande robustesse face à l'ajout d'objets occultants. L'étape de filtrage a été réalisée de manière efficace. Le réseau de neurones EfficientNetv2 a également produit des résultats très satisfaisants. En revanche, le réseau YOLOv5 a montré une variabilité marquée, en générant parfois d'excellents résultats et d'autres fois des résultats moins

$\alpha = 0.05$

	Indépendante	2 par 2		Globale
EfficientNetv2	302	230		104
Méthode proposée	323		143	
YOLOv5	186			

$\alpha = 0.01$

	Indépendante	2 par 2		Globale
EfficientNetv2	230	159		43
Méthode proposée	282		73	
YOLOv5	116			

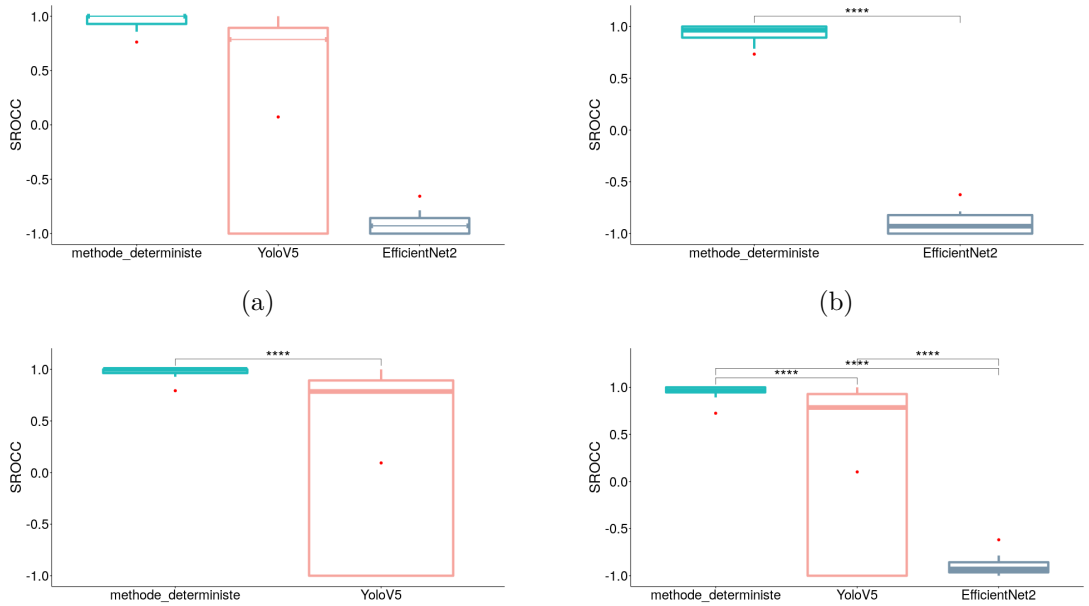
TABLE III.3. – Coefficients de corrélation associés à la dégradation *occultation*. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaison effectuée parmi les 441 classements de références disponibles.

convaincants.

6.8.3. Analyses des performances par rapport à la dégradation *changement d'échelle*

Dans l'analyse indépendante, dont les résultats statistiques sont présentés dans la Figure III.33a, la méthode déterministe proposée se distingue avec une moyenne élevée de 0.76, indiquant une forte cohérence avec les classements de référence.

En revanche, EfficientNetv2 et YOLOv5 présentent des moyennes respectives de -0.66 et 0.072, ce qui met en avant leur difficulté à gérer les changements d'échelle. L'écart interquartile IQR met en évidence la stabilité, avec 0.07 pour EfficientNetv2 mais aussi l'imprévisibilité 1.89 pour YOLOv5. Ainsi, la méthode déterministe proposée démontre une performance plus stable et cohérente face au changement d'échelle. Dans la comparaison suivante, entre le réseau de neurones EfficientNetv2 et la méthode déterministe proposée, cf. Figure III.33b, en conservant les 177 classements significatifs, nous pouvons observer que 75% des classements reconstruits par la méthode déterministe proposée ont un coefficient de corrélation supérieur à 0.83, tandis qu'EfficientNetv2 propose des classements dont 75% ont un coefficient de corrélations inférieur à -0.82. Ensuite, dans la comparaison avec YOLOv5, nous pouvons remarquer que les deux médianes sont satisfaisantes : 1 pour la méthode déterministe proposée et 0.78 pour YOLOv5. Cependant, la méthode déterministe proposée maintient une moyenne élevée de 0.79, alors que YOLOv5 présente une moyenne plus basse de 0.09 ainsi qu'un grand intervalle interquartile IQR de 1.89.



(a) (b)

(c) (d)

FIGURE III.33. – Résultats associés à la dégradation *changement d'échelle*. Nous présentons les performances de chaque méthode en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expériences sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

Enfin, dans la dernière comparaison regroupant les trois méthodes, ce qui ne permet de ne garder que 75 classements significatifs, la méthode déterministe proposée maintient une moyenne élevée de 0.73 alors que YOLOv5 et EfficientNetv2 présentent des moyennes respectives de 0.102 et -0.62. Les médianes associées à YOLOv5 et à la méthode déterministe proposée restent satisfaisantes avec des valeurs respectives de 0.79 et de 0.97, alors que celle de EfficientNetv2 de -0.93 est mauvaise.

Les tendances restent similaires lorsque nous modifions la valeur de seuil pour la *p-value* à $\alpha = 1.10^{-2}$.

Ces résultats illustrent la robustesse aux changements d'échelle ainsi que la supériorité de la méthode déterministe proposée face à YOLOv5 et surtout par rapport EfficientNetv2 qui présente des résultats inattendus, dans le contexte du *changement d'échelle*. En effet, jusqu'à présent le score de confiance associé à la sortir d'EfficientNetv2 obtenaient des résultats exploitables mais, dans ce cas de dégradations, ce n'est plus le cas.

$\alpha = 0.05$

	Indépendante	2 par 2		Globale
EfficientNetv2	243	177		75
Méthode proposée	326		137	
YOLOv5	194			

$\alpha = 0.01$

	Indépendante	2 par 2		Globale
EfficientNetv2	179	110		34
Méthode proposée	292		90	
YOLOv5	137			

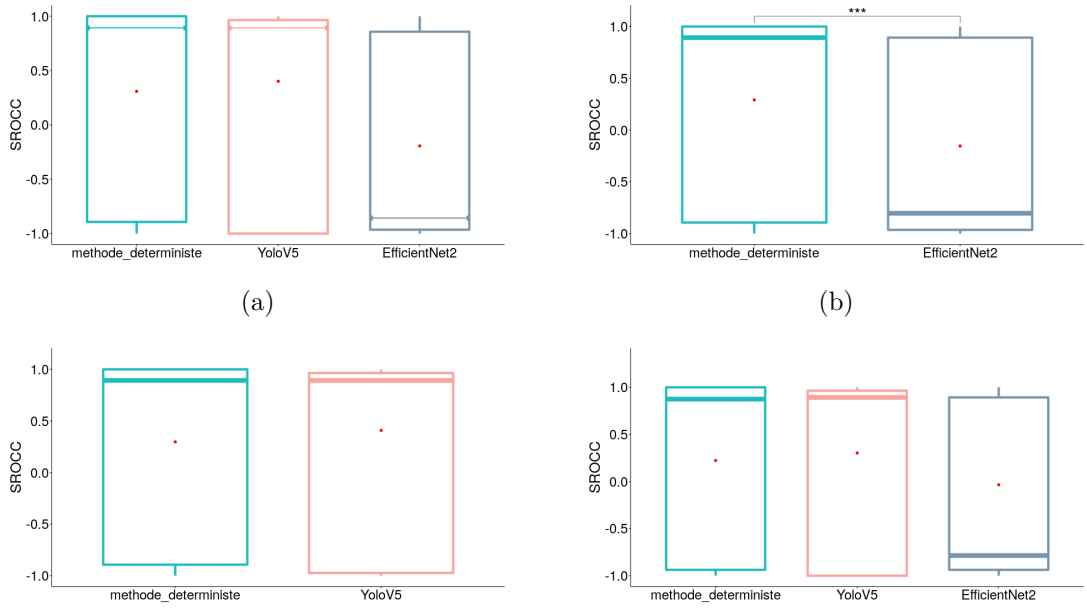
TABLE III.4. – Coefficients de corrélation associés à la dégradation *changement d'échelle*. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaison effectuée parmi les 441 classements de références disponibles.

6.8.4. Analyses des performances par rapport à la dégradation *luminosité*

Nous pouvons remarquer, d'après les résultats statistiques de la Figure III.34a, que les trois méthodes ont un comportement variable face à ce type de dégradation. En effet, les trois méthodes ont un IQR supérieur à 1.8, ce qui montre une grande sensibilité face au changement d'éclairage. Cependant, les méthodes de YOLOv5 et la nôtre montrent une certaine efficacité sur 50% des classements significatifs obtenus, c'est-à-dire sur, respectivement, 126 et 137 classements, cf. la Table III.5. Tandis que pour le réseau EfficientNetv2, 87 classements reconstruits considérés comme significatifs ont obtenu un score de corrélation inférieur à -0.86.

Ces observations restent valables lorsque nous comparons directement la méthode déterministe proposée avec chacun des scores de confiance des deux réseaux, puis de manière simultanée sur les mêmes ensembles de données significatifs. En effet, d'après les diagrammes des Figures III.34c et III.34d, la méthode déterministe proposée et le réseau YOLOv5 possèdent des résultats très similaires, alors que EfficientNetv2 montre des difficultés lorsqu'il y a des changements de luminosité au sein des classements de référence.

La similarité des résultats entre la méthode déterministe proposée et YOLOv5 demeure cohérente, même lorsque nous modifions le seuil α pour la *p-value* fixé à $\alpha = 1.10^{-2}$. Selon les données de la Table III.5, dans le cadre de l'étude indépendante, la méthode déterministe proposée et YOLOv5 préservent environ 80% de leurs coefficients de corrélation significatifs, tandis qu'EfficientNetv2, qui manifestait déjà certaines limitations, voit 32% de ces coefficients ne plus être considérés comme significatifs.



(a) (b)

(c) (d)

FIGURE III.34. – Résultats associés à la dégradation *luminosité*. Nous présentons les performances de chaque méthode en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expériences sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

$\alpha = 0.05$

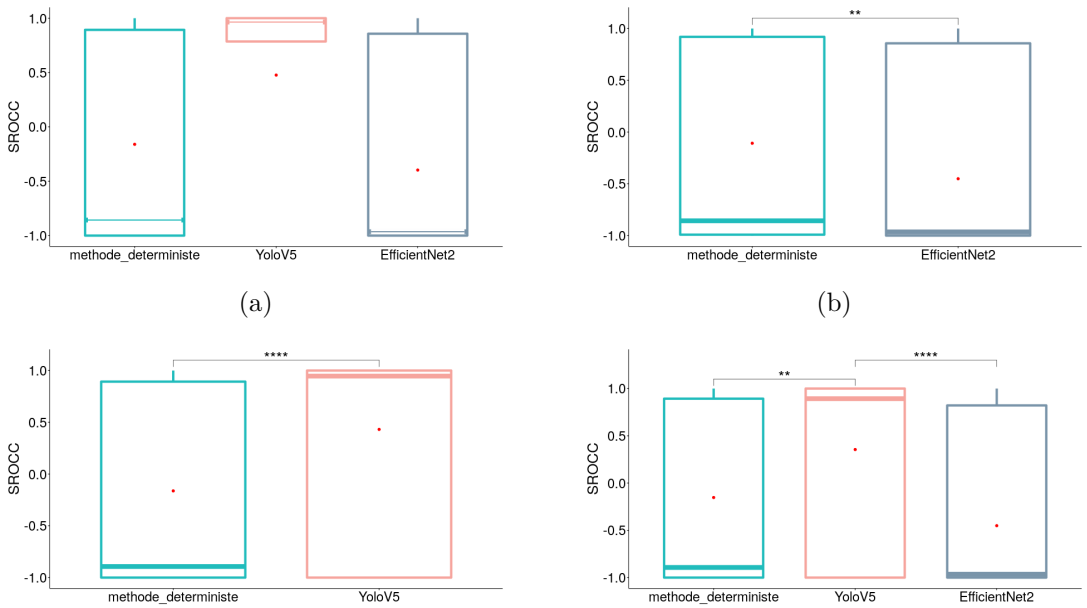
	Indépendante	2 par 2		Globale
EfficientNetv2	194	126		76
Méthode proposée	274		154	
YOLOv5	257			

$\alpha = 0.01$

	Indépendante	2 par 2		Globale
EfficientNetv2	132	63		37
Méthode proposée	229		106	
YOLOv5	207			

TABLE III.5. – Coefficients de corrélation associés à la dégradation *luminosité*. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaison effectuée parmi les 441 classements de références disponibles.

6.8.5. Analyses des performances par rapport à la dégradation *flou gaussien*



(a) (b) (c) (d)

FIGURE III.35. – Résultats associés à la dégradation *flou gaussien*. Nous présentons les performances de chaque méthode en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expériences sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

À l'opposé des autres dégradations étudiées, notre analyse indépendante révèle que le réseau de neurones YOLOv5 se distingue des deux autres méthodes, comme le montre les diagrammes de la Figure III.35a. Avec une médiane de 0.96 parmi les 250 classements significatifs, comme le montre la Table III.6, YOLOv5 surpasse la méthode déterministe proposée ainsi que le réseau EfficientNetv2, qui affichent respectivement des médianes de -0.85, parmi les 243 classements significatifs, et de -0.96, parmi les 268 classements significatifs. L'écart interquartile IQR souligne cette dispersion, avec des valeurs de 1.85 pour EfficientNetv2 et 1.89 pour la méthode déterministe proposée, tandis que YOLOv5 se stabilise à 0.22. Cependant, lors de la comparaison directe entre la méthode déterministe proposée et chaque approche utilisant un score de confiance, une dynamique différente apparaît. En effet, d'après le diagramme de la Figure III.35c, YOLOv5 produit des résultats moins convaincants lorsqu'il est évalué sur un même ensemble de données que la méthode déterministe proposée. Sur les 136 classements significatifs communs, YOLOv5 affiche une moyenne de 0.43 et un IQR qui grimpe à 2. Cette évolution s'explique par l'élimination des classements significatifs pour YOLOv5, ayant des coefficients de corrélation corrects,

lors de l'intersection avec les classements significatifs de la méthode déterministe proposée. Suite à cette réduction du nombre de classements, passant de 250 à 136 pour YOLOv5,

$\alpha = 0.05$

	Indépendante	2 par 2	Globale
EfficientNetv2	268	142	79
Méthode proposée	243	136	
YOLOv5	250		

$\alpha = 0.01$

	Indépendante	2 par 2	Globale
EfficientNetv2	226	96	46
Méthode proposée	196	99	
YOLOv5	220		

TABLE III.6. – Coefficients de corrélation associés à la dégradation *flou gaussien*. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaison effectuée parmi les 441 classements de références disponibles.

seuls les classements significatifs aux coefficients de corrélation extrêmes sont conservés. La méthode déterministe proposée et le réseau EfficientNetv2 présentent des comportements similaires à travers toutes les comparaisons. Dans chaque cas, 50% des classements présentent un coefficient de corrélation inférieur en moyenne à -0.85 et -0.96, respectivement. Bien que la méthode déterministe proposée génère des résultats non satisfaisants, elle surpasse toujours les performances d'EfficientNetv2.

Cette distinction persiste même lorsque nous utilisons un seuil plus restrictif pour α utilisé pour la *p-value* en le fixant à 1.10^{-2}. Ainsi, YOLOv5 conserve toujours 50% de ses classements avec un coefficient de corrélation élevé, tandis que les deux autres méthodes maintiennent 50% de leurs classements avec un coefficient de corrélation inférieur à environ 0.9.

En conclusion, YOLOv5 se distingue lors d'une étude indépendante avec une médiane élevée parmi les classements significatifs mais il présente une moindre performance lorsqu'il est évalué sur un même ensemble de données que la méthode déterministe proposée. Enfin, la méthode déterministe proposée, bien que fournissant des résultats non satisfaisants, surpasse constamment les performances de EfficientNetv2.

6.8.6. Bilan

Les performances des méthodes sont récapitulées dans la Table III.7. En présence d'une augmentation, la méthode déterministe proposée se démarque par sa robustesse, mainte-

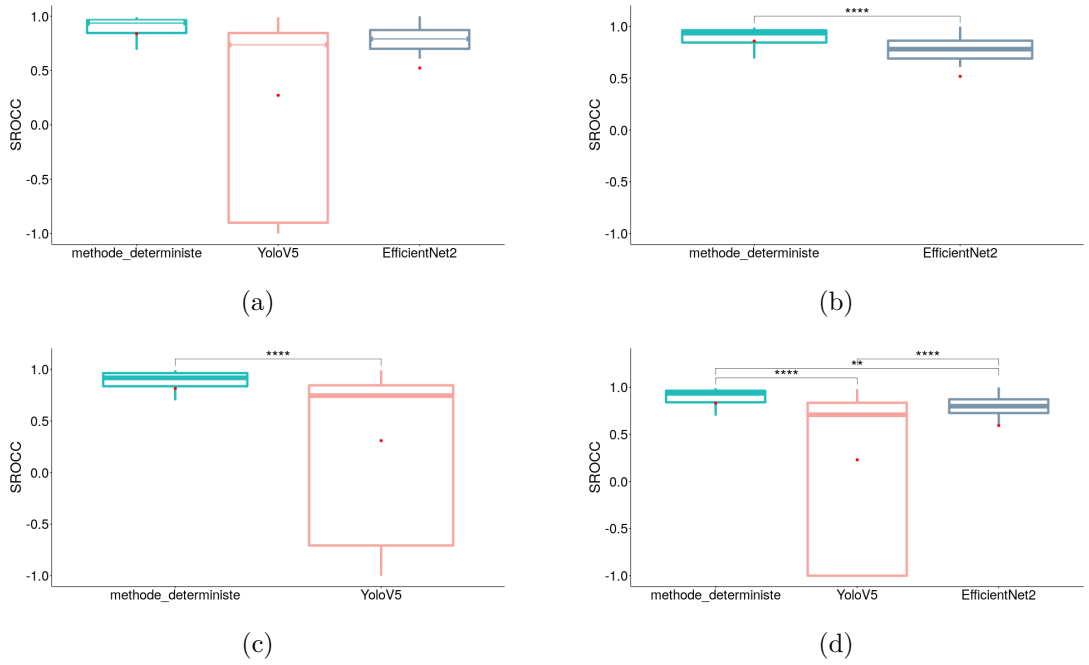
Dégradation	Méthode proposée	EfficientNetv2	YOLOv5
Augmentation	✓	≈	X
Occultation	✓	✓	≈
Changement d'échelle	✓	X	≈
Luminosité	≈	∼	≈
Flou gaussien	∼	∼	≈

TABLE III.7. – Comportement de chaque méthode en fonction des différents types de dégradation : méthode robuste ✓, méthode satisfaisante ≈, méthode imprévisible ∼ et méthode très peu efficace X.

nant des performances élevées, alors qu'EfficientNetv2 présente des résultats moins satisfaisants et YOLOv5 montre une efficacité limitée. Face à l'occultation, la méthode déterministe proposée maintient une robustesse exemplaire, tandis qu'EfficientNetv2 montre une efficacité notable mais perturbée, et YOLOv5 est fortement impacté. Lors du changement d'échelle, la méthode déterministe proposée se distingue en étant à la fois robuste et performante, EfficientNetv2 affiche des performances mitigées, et YOLOv5 présente des résultats variés mais globalement moins satisfaisants. Pour la dégradation par luminosité, les trois méthodes sont affectées, mais la méthode déterministe proposée et YOLOv5 parviennent à fournir des résultats acceptables, contrairement à EfficientNetv2. Enfin, en présence de flou gaussien, la méthode déterministe proposée et EfficientNetv2 montrent une sensibilité avec des résultats peu satisfaisants, tandis que YOLOv5, bien que moins impacté, maintient des performances correctes. En conclusion, en dehors du flou gaussien, la méthode proposée présente le plus de robustesse face aux diverses difficultés étudiées.

6.9. Comparaisons quantitatives sur la combinaison de dégradations

Dans cette étude, les 441 classements de référence considérés contiennent les cinq types de dégradations, avec l'ordre d'application et les paramètres associés détaillés dans la Table III.1, de la section 6.6.2. Tout d'abord, nous pouvons observer que la méthode déterministe proposée génère le plus grand nombre de coefficients de corrélation significatifs, d'après la Table III.8. Plus spécifiquement, le score de pertinence produit des classements significatifs pour 84% des classements de référence, comparé à seulement 59% pour EfficientNetv2 et 56% pour YOLOv5. De plus, d'après les diagrammes de la Figure III.36a, le score de pertinence obtient une moyenne de 0.9 et une médiane de 0.94 surpassent celles



(a) (b) (c) (d)

FIGURE III.36. – Résultats associés à la combinaison de dégradations. Nous présentons les performances de chaque méthode par rapport aux 441 classements de référence, en (a), puis des comparaisons entre la méthode déterministe et chaque approche utilisant un score de confiance : EfficientNetv2, en (b), et YOLOv5 en (c). Nous comparons les trois méthodes étudiées en (d). Les expériences sont réalisées avec un seuil de valeur $p = 5.10^{-2}$.

obtenues par les deux réseaux : la moyenne et la médiane de YOLOv5 sont respectivement de 0.27 et 0.74, tandis que pour EfficientNetv2, elles sont respectivement de 0.52 et 0.79. Cependant, EfficientNetv2 présente des résultats que nous pouvons considérer comme satisfaisants. En effet, tout comme la méthode déterministe proposée, le réseau EfficientNetv2 affiche un IQR bas, indiquant que ces deux méthodes sont prévisibles, contrairement à YOLOv5 qui est imprévisible avec un IQR de 1.74.

Lorsque nous effectuons les deux comparaisons directes entre la méthode déterministe proposée et chacune des approches par score de confiance, les observations demeurent inchangées, comme le montrent les Figures III.36b et III.36c. Dans la Figure III.36b, la méthode déterministe proposée et le réseau EfficientNetv2 restent stables avec une valeur d'IQR de 0.12 pour la méthode déterministe proposée et une valeur d'IQR de 0.17 pour le réseau. Par ailleurs, YOLOv5 reste instable avec une moyenne de 0.31 contre une moyenne de 0.82 pour la méthode déterministe proposée, d'après les diagrammes de la Figure III.36c.

Ensuite, lors de la comparaison simultanée des trois méthodes, la méthode déterministe proposée demeure la plus efficace. En effet, 75% des classements qu'elle obtient ont un coefficient de corrélation supérieur à 0.84, contre 0.73 pour le réseau EfficientNetv2.

Enfin, lorsque nous resserrons notre sélection de classements significatifs, la méthode dé-

$\alpha = 0.05$

	Indépendante	2 par 2		Globale
EfficientNetv2	261	219		127
Méthode proposée	370		206	
YOLOv5	248			

$\alpha = 0.01$

	Indépendante	2 par 2		Globale
EfficientNetv2	214	173		81
Méthode proposée	353		152	
YOLOv5	195			

TABLE III.8. – Coefficients de corrélation associés à la combinaison de dégradations. Nous présentons le nombre de coefficients de corrélation significatifs pour chaque type de comparaison effectuée parmi les 163 classements de références disponibles. Toutes les dégradations ont été prises en compte.

terministe fournit toujours les résultats les plus performants, surpassant les deux réseaux. EfficientNetv2 affiche des résultats légèrement moins satisfaisants, avec une moyenne de 0.55 et une médiane de 0.81, comparé à la méthode déterministe proposée qui a, en général, une moyenne autour de 0.85 et une médiane proche de 0.93. En revanche, le réseau YOLOv5 continue de montrer des performances peu satisfaisantes.

En conclusion, l'évaluation des performances des trois méthodes étudiées, à savoir la méthode déterministe s'appuyant sur le score de pertinence, et les deux approches utilisant un score de confiance issus de EfficientNetv2, et YOLOv5, face à des images subissant successivement différentes dégradations, confirme les résultats obtenus avec chaque dégradation séparément. la La méthode déterministe proposée est efficace et robuste à de nombreuses dégradations. En effet, elle se distingue en produisant un nombre significativement plus élevé de classements pertinents par rapport à EfficientNetv2 et YOLOv5. Les indicateurs tels que la moyenne et la médiane des coefficients de corrélation confirment les performances de cette méthode déterministe, avec des valeurs plus élevées et moins de dispersion, démontrant sa robustesse dans des conditions de dégradations variées. Cependant, EfficientNetv2 présente des résultats satisfaisants, tandis que YOLOv5 montre des performances limitées. Lors des comparaisons directes entre les méthodes, l'approche déterministe reste la plus efficace, et même lorsque les critères de sélection sont rendus plus stricts, elle maintient sa supériorité, soulignant sa stabilité face à des conditions plus contraignantes. Ces observations renforcent l'argument en faveur de l'efficacité de la méthode déterministe proposée dans la génération de classements pertinents dans des

scénarios réalistes de dégradations d'images.

Au cours des expériences précédemment détaillées, nous avons constaté que le réseau de neurones YOLOv5 n'a pas produit de résultats satisfaisants, ce qui était surprenant étant donné la réputation de ce réseau. Pour mieux comprendre son comportement, nous avons analysé ses performances et partagé nos observations dans [Pelissier-Combescure 23]. Plus précisément, dans ces travaux, les classements de référence que nous avons construits s'appuyaient sur uniquement trois types de dégradations : l'*augmentation*, l'*occultation* et le *changement d'échelle*, répétés trois fois chacun. Nous avons remarqué que YOLOv5 n'était pas assez fiable pour fournir, en permanence, des scores de confiance. Après analyse, nous avons observé que, en moyenne, sur les dix images de chaque classement significatif, YOLOv5 parvenait à détecter l'objet étudié et à fournir un score de confiance pour seulement sept images, tandis que le réseau EfficientNetv2 réussissait à classer en moyenne neuf images. Nous avons noté que les images, pour lesquelles les réseaux ne pouvaient pas fournir de score de confiance, étaient soit de résolution insuffisante, soit comportaient trop d'occultations. Cette observation pourrait expliquer les performances médiocres de YOLOv5 sur certains des classements de vérité terrain. De plus, ces observations sont cohérentes avec les résultats présentés dans les § 6.8.2 et 6.8.3, où nous avons constaté que les deux approches basées score de confiance étaient inefficaces face à la dégradation *changement d'échelle*, et en ce qui concerne YOLOv5, que ses résultats sur les classements contenant uniquement des *occultations* étaient peu satisfaisants. Cependant, ces classements n'ont pas été exclus de notre processus d'évaluation car cela fausserait la vérité sur les performances réelles des scores de confiance et les favoriserait nettement. Une autre explication plausible à ces résultats inattendus pourrait être que les réseaux ont été principalement entraînés sur des images avec un contexte, c'est-à-dire composées de plusieurs objets et de textures différentes. Ils ont donc appris à extraire des caractéristiques de l'ensemble de la scène, et leurs performances diminuent lorsque l'objet est hors contexte, ce qui correspond à un environnement peu fréquent dans les images d'entraînement.

6.10. Résultats qualitatifs

Les analyses décrites précédemment ont montré la capacité de la méthode déterministe proposée à retrouver l'ordre correct pour des images dégradées de manière synthétique. Cependant, la finalité de la méthode déterministe est de sélectionner des images pour le visuel qu'elles proposent. Il est donc nécessaire de vérifier visuellement si les résultats sur des images réelles qui n'ont pas été dégradées manuellement sont aussi satisfaisants. Encore une fois, nous nous appuyons sur la base de données Pix3D [Sun 18].

La Figure III.37 affiche les classements, pour deux ensembles de quatre images contenant le même canapé, issue des trois méthodes étudiées et comparées. Pour le premier

ensemble de la Figure III.37a, la méthode déterministe proposée est la seule à positionner l'image du canapé blanc en première place. Cette image est par définition l'image avec la meilleure mise en valeur : l'objet est dominant, de grande taille et il n'y a aucune occultation possible, car l'environnement est vide. Cependant, l'image avec le canapé vert offre également une mise en valeur satisfaisante. La seule différence concerne l'environnement, qui est ici plus complexe. C'est pour cette raison que nous pouvons affirmer que l'approche utilisant le score de confiance issu du réseau de neurones EfficientNetv2 apporte des résultats satisfaisants, mais est moins efficace que la méthode déterministe. Contrairement à YOLOv5 qui place en deuxième position une image dans laquelle la visibilité du canapé est réduite à cause de son orientation et surtout à cause de la présence d'occultations (coussins et table basse). Par ailleurs, la Figure III.37b présente d'autres résultats issus d'un second ensemble d'images. Une fois de plus, la méthode déterministe proposée se démarque en produisant le classement le plus pertinent, plaçant en première position l'image mettant en évidence le canapé rouge de manière dominante et sans aucun contexte autour. Nous pouvons remarquer que les images suivantes sont arrangées en fonction de la surface occultée du canapé, formant ainsi une séquence logique. Ce classement diffère de ceux générés par les deux approches utilisant des scores de confiance. Plus précisément, les classements ne suivent pas une tendance croissante de la surface occultée. Notamment, YOLOv5 positionne l'image présentant un canapé blanc, contenant un grand nombre d'objets occultants, avant celle du canapé rouge, où aucun objet occultant n'est présent. Cette observation souligne la vulnérabilité du réseau face aux occultations, un aspect déjà identifié lors de notre analyse quantitative détaillée dans le § 6.8.2. À ce stade de nos évaluations et de notre analyse, il est important de rappeler que mesurer la pertinence d'une image par rapport à un objet qu'elle contient n'est pas la tâche initiale à laquelle les réseaux ont été entraînés. Ce détail peut expliquer les résultats non convaincants obtenus au niveau de score de corrélation ou au niveau des résultats visuels.

Pour approfondir et confirmer nos observations, nous appliquons l'approche déterministe à un ensemble d'images brutes contenant le même objet afin de les classer en fonction de leur capacité à mettre en valeur l'objet étudié. Nous avons testé l'approche déterministe sur 9 modèles de la catégorie *Sofa*, soit 1048 images. Les exemples de classement sont présentés dans les Figures III.18 et III.38. Nous remarquons que les dernières images mettent le moins en valeur l'objet parce qu'il est tronqué ou mal positionné. Plus précisément, les dernières images ont les plus faibles scores de pertinence en raison de la présence d'objets occultants, d'une faible information caractéristique et d'une faible dominance. En revanche, les images qui présente l'objet d'intérêt sans environnement ou où l'objet est dominant avec très peu d'occultations sont toujours placées en tête du classement. Ces exemples illustrent la robustesse et l'efficacité de la méthode déterministe proposée.

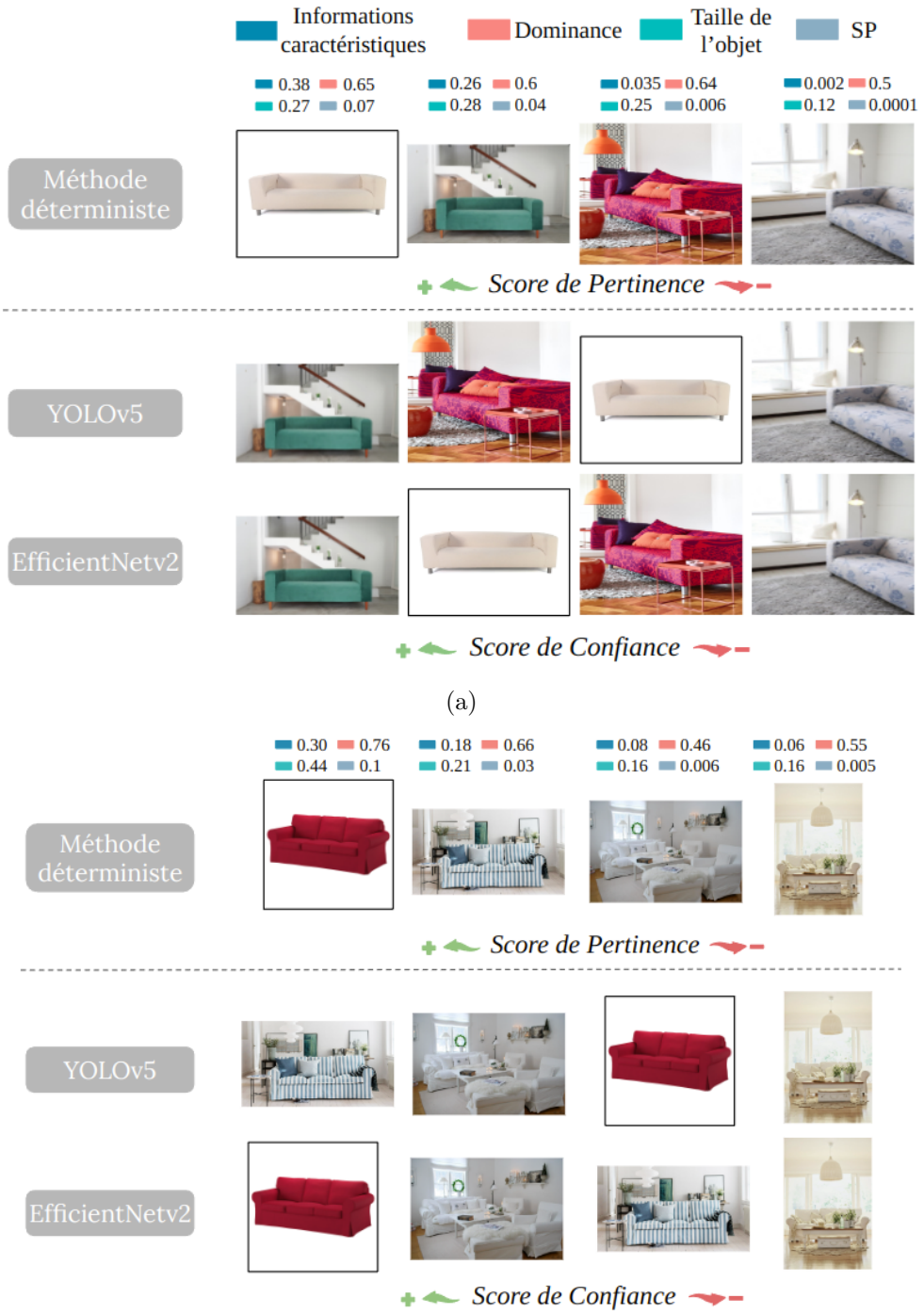


FIGURE III.37. – Comparaisons visuelles entre le score de pertinence et le score de confiance. Classements d'images en fonction soit du score de pertinence proposé, soit du score de confiance, obtenus avec les trois méthodes étudiées sur des images réelles contenant le même objet.

FIGURE III.38. – Classements d'images réelles obtenus avec l'approche déterministe. Nous présenton le détail des valeurs obtenus pour chacun des trois terme du score de pertinence.

En résumé

Étant donné un objet 3D, notre objectif est de déterminer, parmi un ensemble images 2D, celles qui représentent le mieux cet objet, et de les classer. Les images sélectionnées doivent être à la fois informatives et offrir une vue pertinente de l'objet, c'est-à-dire des visuels qui présentent le plus d'information essentielle possible à propose de l'objet 3D. Pour estimer la qualité de la vue, nous proposons de nous appuyer sur des points répétables de second ordre, extraits à l'aide d'un détecteur de saillance curviligne, afin de calculer un score de pertinence, indépendant des couleurs et des textures. Ce score se compose de trois termes reposant sur des concepts photographiques : la quantité d'information caractéristique disponible, la place occupée par l'objet et la taille de l'objet par rapport au reste de l'image. Sur la base de ce score, et compte tenu d'un ensemble d'images contenant le même objet, nous sommes en mesure de classer ces images, de la plus pertinente à la moins pertinente.

Les réseaux de neurones dédiés à la détection et à la classification sont capables de reconnaître des objets avec un score de confiance, et certains travaux dans la littérature ont déjà mis en compétition ces derniers avec des êtres humains sur des tâches subjectives. Nous nous sommes inspirées de ces travaux pour développer une approche automatique s'appuyant sur ce score de confiance extrait des réseaux de neurones.

Pour évaluer et comparer la méthode déterministe proposée et les méthodes exploitant un score de confiance, nous avons construit nos propres jeux de données afin d'avoir une vérité terrain. Ces classements objectifs de référence sont composés d'images ayant subi successivement diverses dégradations simulées. Nous avons utilisé le coefficient de corrélation de Spearman pour évaluer la pertinence des classements obtenus par les trois méthodes étudiées et comparées avec les classements attendus. Nous fournissons également des résultats qualitatifs visuels sur un ensemble de données réelles. Les résultats montrent sans équivoque l'efficacité et la robustesse de la méthode déterministe proposée. De plus, ces résultats aident à comprendre le comportement des méthodes s'appuyant sur un score de confiance.

Ces travaux ont été présentés, en 2022, lors du GTMG (*Groupe de Travail en Modélisation Géométrique*) et par un poster aux JNIM (*Journées Nationales de l'Informatique Mathématique*), en 2023. Ils ont également été publiés lors de la conférence nationale : RFIAP (*Reconnaissance des Formes, Images, Apprentissage et Perception*) [Pelissier-Combescure 22], en 2022, et lors de la conférence internationale SCIA (*Scandinavian Conference on Image Analysis*) [Pelissier-Combescure 23] en 2023.

Chapitre IV

Points de vue non restreints

Sommaire

7. Mesure de la pertinence d'une vue d'un objet 3D . . . 125
   7.1. Introduction . . . 125
   7.2. Extraction et combinaison d'attributs géométriques . . . 128
   7.3. Méthodes de saillance 3D et formules d'angle testées . . . 131
      7.3.1. Saillance intrinsèque 3D . . . 131
      7.3.2. Détection des faces et sommets visibles . . . 132
      7.3.3. Pondération de la saillance . . . 133
   7.4. Proposition de validation . . . 134
      7.4.1. Classification des techniques de validation . . . 134
      7.4.2. Base de données . . . 134
      7.4.3. Étude utilisateurs et utilisatrices . . . 134
      7.4.4. Score de proximité . . . 137
   7.5. Optimisation de notre approche . . . 140
      7.5.1. Choix de la méthode de saillance 3D . . . 140
      7.5.2. Étude d'ablation . . . 141
   7.6. Comparaisons quantitatives et qualitatives . . . 141
8. Réalisation d'une étude utilisateurs et utilisatrices . . . 146
   8.1. Contexte . . . 146
   8.2. Présentation de l'interface . . . 149
      8.2.1. Page d'introduction . . . 149
      8.2.2. Définition du « bon » point de vue . . . 149
      8.2.3. Tutoriel proposé et consignes . . . 150
      8.2.4. Justification des choix des utilisateurs . . . 151
      8.2.5. Stockage des données . . . 151
   8.3. Fonctionnalités de l'interface . . . 151
   8.4. Pré-traitement des caméras et des modèles 3D . . . 154
   8.5. Post-traitement des données récoltées . . . 156
      8.5.1. Réception des données . . . 156

8.5.2. Extraction d'informations statistiques 158
8.5.3. Filtrage des participants 159
8.5.4. Construction des histogrammes de popularité des points de vue 161
8.6. Analyses et interprétations des histogrammes 162
8.6.1. Cas général 162
8.6.2. Vues accidentelles 164
8.6.3. Vues occultées 165
8.6.4. Objets avec des yeux 165
8.6.5. Objets symétriques 166
8.6.6. Objets non-familiers 166
8.7. Comparaisons avec des images réelles 168

Problématique

Dans ce chapitre, nous abordons l'évaluation de la pertinence d'un point de vue en s'appuyant directement sur la représentation de notre objet d'intérêt par un maillage surfacique 3D. Contrairement au chapitre précédent où les points de vue disponibles de l'objet étudié étaient uniquement ceux présents dans une collection d'images donnée, à présent, nous sommes libérées de cette contrainte. En effet, en travaillant à partir des modèles 3D, nous avons accès à la totalité des points de vue disponibles et envisageables. Les modèles 3D sont intrinsèquement indépendants du point de vue, offrant ainsi la possibilité d'extraire des informations géométriques à partir de n'importe quel angle de vue. Ce n'est pas le cas avec les images qui sont étroitement liée à un point de vue. Néanmoins, l'objectif ici demeure similaire à celui du chapitre précédent : évaluer la pertinence de chacune des vues d'un objet 3D, en utilisant, dans ce chapitre, son maillage 3D.

Ainsi, nous explorons une nouvelle manière de calculer la pertinence d'une vue par rapport à un objet d'intérêt dans la section 7. Il s'agit d'une approche géométrique. Puis, dans la section 8, nous introduisons une étude utilisatrice et utilisateur pour valider la cohérence des vues sélectionnées par l'approche géométrique proposée. Plus précisément, nous souhaitons montrer que la sélection automatique qui est réalisée est proche de celle qui pourrait être faite par un être humain.

7. Mesure de la pertinence d'une vue d'un objet 3D

7.1. Introduction

Cette section vise à présenter une méthode permettant de mesurer la pertinence des vues d'un objet afin de sélectionner la plus adéquate. Plus globalement, notre travail s'inscrit dans une problématique plus large : comment visualiser un objet 3D à travers un support en 2D ? Effectivement, dans notre quotidien, nous sommes entourés de représentation 2D d'objets 3D : les affiches publicitaires, les dessins, les tableaux d'art, les photographies ou encore les rapports illustrées avec des images. Cela n'est pas sans raison, car bien souvent une représentation 2D est plus parlante et porteuse de sens qu'un long texte. Cette tâche de sélection d'un point de vue s'inscrit dans le domaine de la recherche en sélection du meilleur point de vue, *Best View Selection*, que nous pouvons définir comme la quête du point de vue optimal pour visualiser de manière idéale un objet 3D donné.

Objectif

Déterminer le point de vue le plus pertinent d'un objet 3D, c'est-à-dire permettant sa compréhension globale sans ambiguïté et utilisant uniquement une représentation par un maillage 3D.

Dans nos travaux, nous avons considéré qu'une bonne vue 2D d'un objet 3D était celle qui qui permet de l'identifier sans ambiguïté. Comme dans le chapitre précédent, nous nous sommes appuyées sur l'extraction d'information essentielle disponible dans chacune vue étudiée. Nous recherchons celles qui contiennent le plus d'information et d'éléments caractéristiques de cet objet. En effet, avec un point de vue pertinent, identifier la nature d'un objet devient une tâche facile pour une personne. Il est d'autant plus confortable et agréable de pouvoir reconnaître un objet en un seul coup d'œil que de devoir le manipuler dans tous les sens pour le distinguer. Cela permet également un gain de temps et d'efficacité.

Pour illustrer et justifier ces avantages, nous citons le rapport de [Fabricatore 02]. Ce dernier fournit la liste des éléments clés à prendre en compte pour créer un « bon » jeu vidéo du point de vue des joueuses et des joueurs. Ces éléments peuvent concerner la jouabilité, l'interface entre le jeu et les personnes qui l'utilisent, le scénario ou encore le décor. Plus précisément, ce rapport distingue deux catégories d'informations transmises aux joueurs et joueuses durant une partie : les informations esthétiques et les informations fonctionnelles. Les premières englobent la majorité des éléments du contexte du jeu et ont pour objectif principal de créer une ambiance immersive. Elles visent à capter l'attention sur un plan émotionnel en offrant la sensation de participer à un univers virtuel attrayant. La seconde catégorie d'information est essentielle afin que la joueuse ou le joueur accomplisse les actions nécessaires pour progresser dans le jeu. Si nous considérons ce deuxième type d'information, il est important de choisir des points de vue pertinents, et pas nécessairement esthétiques, pour les objets qui permettent d'avancer dans le jeu. Un point de vue qui permet une identification rapide d'un objet, et par conséquent la compréhension de sa fonctionnalité, est indispensable pour que les joueurs et joueuses comprennent l'action à réaliser et ainsi avoir un déroulement optimal et agréable du jeu. Voici un extrait du rapport qui appuie cette idée : « *The player must be able to clearly understand the semantics of objects in order to interact with them.* »

Une manière de déterminer ces bons points de vue serait d'utiliser une technique par apprentissage. En effet, il existe des centaines de bases de données contenant des dizaines de milliers d'images. Cependant, il existe deux biais qui posent un problème vis-à-vis de notre problématique : la mise en avant de l'esthétisme et la limitation des points de vue accessibles via les images disponibles. La Figure IV.1 illustre le premier biais lié à

(a) (b) (c)

FIGURE IV.1. – Illustration du biais lié à l'esthétisme avec l'exemple d'une tasse. Nous présentons trois images d'une tasse : en (a), avec un point de vue pertinent qui peut être également considéré comme esthétique, en (b), avec un point de vue ambigu sur la hauteur et, en (c), avec un point de vue où la tasse peut être confondue avec un bol.

l'esthétisme en présentant le cas de la visualisation d'une tasse. La vue offerte par l'image de gauche, Figure IV.1a, est une vue pertinente de la tasse. En effet, la anse de la tasse, qui est un des éléments caractéristiques, est bien visible et identifiable. Avec une vue légèrement de dessus, sa fonctionnalité, qui est d'être un contenant, est facilement compréhensible. De plus, cette vue peut également être considérée comme esthétique, ces deux propriétés ne sont pas incompatibles. Cependant, dans l'image au centre de la Figure IV.1b, l'esthétisme a été favorisé et cela a créé une ambiguïté sur la profondeur de la tasse. Il est difficile d'avoir une compréhension globale de la tasse et d'appréhender sa taille. Enfin, dans l'image de droite, Figure IV.1c, la anse n'est plus visible, ce qui engendre la confusion entre une tasse et un bol.

(a) (b) (c)

FIGURE IV.2. – Illustration du biais de variabilité limitée des points de vue accessibles avec l'exemple de l'avion. En (a) et (b), ce sont des points de vue de dessous et, en (c), un point de vue à hauteur d'yeux. Celui-ci sera plus fréquemment présent dans les images disponibles que les points de vue (a) et (b).

En ce qui concerne le second biais lié à la variabilité limitée des points de vue accessibles, il faut considérer les objets dont certains points de vue sont difficiles d'accès, par exemple, en raison de leur taille. En effet, il est peu pratique de prendre en photo un avion vu de haut. Grâce à leurs grandes tailles et à leur fonctionnalité, les images des avions sont

souvent vues de dessous ou à hauteur d'yeux, comme l'illustre la Figure IV.2. Les vues de dessus pour les avions sont relativement rares. Il y a donc une inégalité dans la répartition des points de vue, ce qui est un problème pour effectuer l'apprentissage.

En conclusion, d'après ces observations, l'utilisation de techniques par apprentissage ne semble pas complètement adaptée à notre problématique. Nous avons donc décidé d'élaborer une méthode automatique de sélection de meilleure vue en utilisant uniquement l'information fournie par les maillages 3D non texturés des objets. Ainsi, nous pouvons éviter les deux biais cités. En effet, les maillages 3D étant sans contexte et sans texture, contrairement aux images, le biais esthétique est limité. De plus, toutes les vues sont par définition accessibles, et ainsi nous évitons le biais lié à la variabilité limitée des points de vue accessibles. Enfin, il est important de souligner que nous avons également pris en compte certaines préférences humaines au travers de propriétés sémantiques propres à certains types d'objets manipulés.

Nous venons de faire le bilan des différents défis et enjeux de la sélection des vues les plus pertinentes pour des objets 3D. La méthode que nous avons élaborée est détaillée dans le § 7.2, en présentant en détails le processus de sélection et en expliquant le rôle et l'interprétation de chaque concept utilisé. Ensuite, nous présentons les différentes méthodes de saillance et formules d'angle testées, cf. § 7.3, pour évaluer leur impact sur la pertinence des vues sélectionnées automatiquement. La section 7.4 expose notre stratégie pour évaluer la pertinence des résultats obtenus et en particulier nous introduisons l'étude utilisateur et utilisateur qui nous a permis d'obtenir des données représentatives des préférences humaines formant ainsi la vérité terrain. En effet, notre approche vise à identifier les points de vue porteurs de sens pour les êtres humains, car ce sont les bénéficiaires envisagés. Ensuite, la section § 7.5 nous permettra de présenter les choix optimaux des paramètres pour la méthode proposée. Enfin, la section 7.6 est dédiée à la présentation des résultats obtenus et à la comparaison avec les approches de l'état de l'art.

7.2. Sélection du point de vue optimal par extraction et combinaison d'attributs géométriques

Notre approche pour sélectionner automatiquement le meilleur point de vue d'un objet 3D donné, est illustrée dans la Figure IV.3 et s'appuie principalement sur l'extraction de l'information essentielle de l'objet 3D étudié. Plus précisément, nous prenons en compte la saillance intrinsèque de l'objet 3D relative à chaque un point de vue. Il est donc nécessaire de déterminer la visibilité de chaque sommet, comme dans [Lee 05] ou de chaque face, comme dans [Plemenos 04]. Cependant, pour une vue donnée, tous les sommets visibles n'ont pas la même importance, et ne font pas tous partie de l'information essentielle que nous souhaitons extraire. En effet, par définition, l'information essentielle est un sous-ensemble plus compact, constitué des éléments les plus caractéristiques disponibles dans

chaque vue. Pour effectuer cette sélection de points saillants, nous pondérons la saillance des sommets visibles par leur angle de vue par rapport à la position de la caméra.

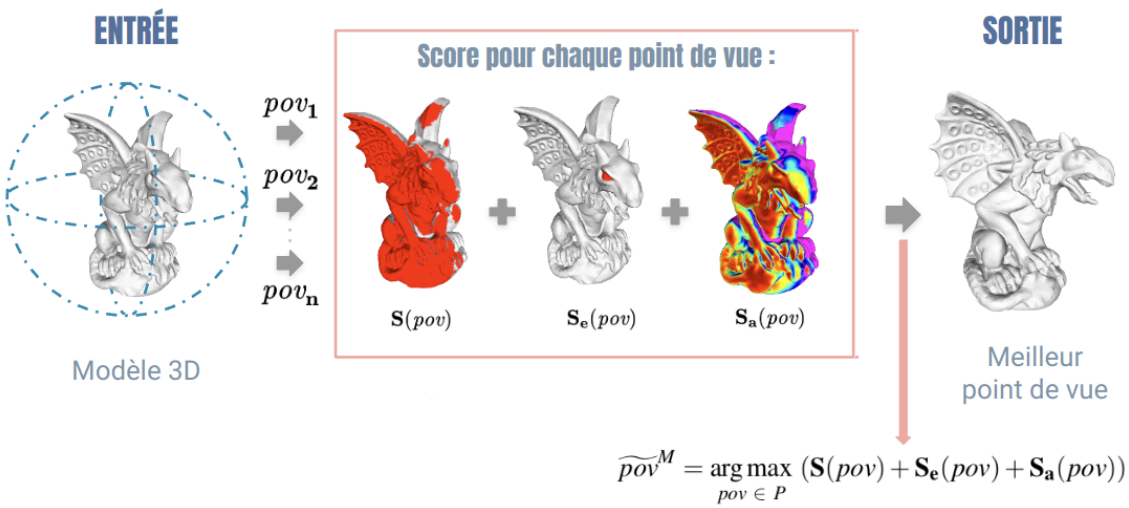


FIGURE IV.3. – Chaîne de traitement pour sélectionner le meilleur point de vue d'un objet 3D. À partir d'un modèle 3D, et pour chaque vue étudiée, nous calculons le pourcentage de surface visible, le pourcentage de surface visible liée aux yeux lorsque cela fait du sens, ainsi que la somme pondérée de la saillance intrinsèque visible. Le point de vue qui maximise la somme de ces trois valeurs est considéré comme le meilleur point de vue de l'objet 3D.

L'approche que nous proposons attribue à chaque vue un score relatif à la pertinence de la partie visible de l'objet. Ainsi, la meilleure vue est celle qui expose le plus possible la complexité géométrique de l'objet. En nous inspirant de nombreux travaux de la littérature, nous avons défini trois termes permettant de regrouper toutes les caractéristiques nécessaires pour mesurer la pertinence d'une vue avec pour objectif de tendre vers une réponse en lien avec la logique humaine. Plus précisément, étant donné un modèle 3D M et un point de vue pov, nous avons introduit :

- la **surface visible** $\mathbf{S}(pov)$: ce terme permet de quantifier la proportion de surface visible en fonction du point de vue pov, c'est-à-dire le rapport entre l'aire 3D visible et l'aire 3D totale du modèle. Ce terme est défini par :

$$\mathbf{S}(pov) = \frac{\mathcal{A}_{3D} \text{ surface visible}}{\mathcal{A}_{3D} \text{ surface totale}} \tag{IV.1}$$

  Ce terme est inspiré des travaux détaillés dans [Secord 11, Kwon 20]. En effet, d'après ces travaux, l'attribut le plus important est celui qui considère la visibilité des surfaces [Secord 11]. Cet attribut fait également partie des quatre meilleures mesures à utiliser selon les observations réalisées dans [Dutagaci 10] pour estimer la vue opti-

male d'un objet 3D.

- la **surface des yeux visible $\mathbf{S_e}(pov)$** : ce terme représente la proportion de surface visible relative à un point de vue, en se concentrant sur la zone des yeux. Autrement dit, le rapport entre l'aire 3D visible des yeux et l'aire 3D totale des yeux. Ce terme est donc la spécialisation aux yeux du terme précédent $\mathbf{S}(pov)$ et est défini par :

$$\mathbf{S}(pov) = \frac{\mathcal{A}_{3D} \text{ surface visible des yeux}}{\mathcal{A}_{3D} \text{ surface totale des yeux}} \tag{IV.2}$$

  En effet, lorsqu'un modèle représente une créature dotée d'yeux ou d'un visage, il a été établi dans [Zusne 70] que les êtres humains montrent une préférence pour les points de vue où les yeux sont visibles. Nous pouvons également justifier l'utilisation de cette information sémantique avec les observations des études utilisateurs et utilisatrices menées dans [Chen 12, Jiang 15]. Ces recherches montrent que les zones du visage, notamment les yeux, attirent toujours un fort intérêt visuel. Ces observations sont également partagées avec les auteurs de [Lavoué 18], qui mentionnent que pour les objets avec un visage, les yeux concentrent une forte attention. Enfin, une autre justification est proposée dans [Secord 11] : le poids associé à l'attribut relatif à la présence des yeux dans la forme linéaire de leur formule est le plus élevé. Pour tenir compte de ces observations, dans nos travaux, nous avons annoté manuellement les faces des modèles 3D correspondant aux yeux. Comme pour les méthodes automatiques de détection des visages et des yeux utilisées dans les images, par exemple [More 21], nous anticipons que des algorithmes similaires pour les modèles 3D non texturés deviendront des outils robustes, éliminant ainsi le besoin de ce processus manuel.

- la **saillance des sommets visibles $\mathbf{S_a}(pov)$** : ce terme quantifie la quantité de saillance intrinsèque 3D visible selon le point de vue (pov). L'utilisation d'un terme relatif à la saillance intrinsèque possède plusieurs justifications. Plus généralement, nous souhaitons pouvoir sélectionner la vue qui permet l'identification de l'objet sans ambiguïté. Pour cela, les auteurs de [Song 19] mentionnent que les informations essentielles à la classification des objets sont importantes pour la saillance, car elles peuvent aider les êtres humains à reconnaître rapidement un objet sans avoir besoin d'en explorer tous les détails. De plus, dans les bilans réalisés dans [Dutagaci 10], respectivement [Secord 11], l'attribut relatif à la saillance fait partie des deux, respectivement des quatre, meilleurs attributs à utiliser. Cette considération de la saillance en fonction d'une vue a été faite de différentes manières dans les approches existantes. Par exemple, dans [Lee 05], les auteurs somment simplement les saillances des sommets visibles. L'implication de la saillance d'un sommet est donc

simplement binaire. Les auteurs de [Feixas 09] et [Leifman 16] ont utilisé une prise en compte de la dépendance au point de vue plus élaborée en additionnant les saillances des sommets visibles, pondérées par une fonction f dépendant de l'angle α_v entre les sommets visibles et la caméra. Nous nous sommes inspirées de ces derniers travaux pour proposer le terme suivant :

$$\mathbf{S_a}(pov) = \sum_{v \in V} S_i(v).f(\alpha_v), \tag{IV.3}$$

avec $S_i(v)$ la saillance intrinsèque du sommet v et V l'ensemble des sommets visibles à partir de pov. Nous avons testé cette formule en considérant différentes méthodes pour $f(\alpha_v)$ et $S_i(v)$. Toutes les possibilités sont détaillées dans la section 7.3. Ce terme de saillance est normalisé en le divisant par la valeur maximale obtenue pour tous les points de vue étudiés.

Une fois que ces trois termes ont été calculés sur l'ensemble P des vues étudiées, nous pouvons déterminer la meilleure vue $\widetilde{pov}^M$ du modèle 3D M. Autrement dit, celle qui maximise la somme des trois termes :

$$\widetilde{pov}^M = \underset{pov \ \in \ P}{\arg\max} \ (\mathbf{S}(pov) + \mathbf{S_e}(pov) + \mathbf{S_a}(pov)) \tag{IV.4}$$

Cette combinaison de termes est inspirée par les résultats obtenus dans [Polonsky 05, Secord 11, Rudoy 12, Kwon 20, Marsaglia 21], qui ont démontré l'efficacité de l'utilisation conjointe de plusieurs attributs géométriques dans la sélection des meilleures vues.

7.3. Méthodes de saillance 3D et formules d'angle testées

7.3.1. Saillance intrinsèque 3D

La saillance des sommets peut être utilisée pour mettre en évidence les zones d'intérêt et diriger l'attention visuelle vers les caractéristiques les plus importantes d'un modèle 3D. Comme nous l'avons montré dans notre état de l'art de la section 4, il existe un grand nombre de méthodes différentes pour estimer la saillance intrinsèques de maillages 3D. Dans nos travaux, nous nous sommes concentrées sur les caractéristiques saillantes calculées suivant les techniques les plus populaires : les courbures moyennes [Lee 05], l'analyse spectrale [Song 14], la classification [Tasse 15], la notion de distinction [1] [Leifman 16] et l'entropie [Limper 16].

1. La distinction d'un sommet est une information locale qui indique à quel point un sommet diffère de ses voisins.

7.3.2. Détection des faces et sommets visibles

Pour connaître quels sommets sont visibles, nous avons d'abord déterminé les faces visibles. Étant donné un point de vue *pov*, nous déterminons quelles faces font face à la caméra en utilisant le principe du *back face culling* [Back-face culling 24]. Plus précisément, une face F est orientée vers la caméra, si le cosinus de l'angle α_F entre sa normale sortante $\overrightarrow{n_F}$ et le vecteur de la caméra $\overrightarrow{pov - c_F}$, avec c_F le centre de F, est supérieur à $\epsilon = 10^{-5}$. Cependant, certaines de ces faces qui sont correctement orientées par rapport à la caméra, peuvent être occultées. Étant donné une face F, $c_F = (x_{c_F}, y_{c_F})$ sont les coordonnées 2D de son centre et z_{c_F} sa profondeur après la projection 2D. Pour les filtrer, nous utilisons les informations de profondeur contenues dans les cartes de profondeur disponibles pour chaque point de vue. Une face est considérée comme visible si la profondeur z_{c_F} associée à la projection 2D de son centre (nous prenons le barycentre) est identique à celle contenue dans la carte de profondeur z_D.

Pour déterminer la profondeur exacte se trouvant dans la carte de profondeur à la position du centre, nous effectuons trois interpolations :

- z_b est déterminé par interpolation entre (i_b, j_b) et (i_a, j_b),
- z_a est déterminé par interpolation entre (i_b, j_a) et (i_a, j_a),
- z_D est déterminé par interpolation entre z_b et z_a.

Si $|z_D - z_{c_F}| \leq 10^{-2}$, alors F est visible. Ensuite, la profondeur z_D associée à (x_{c_F}, y_{c_F}) dans la carte de profondeur D est calculée par interpolation bilinéaire.

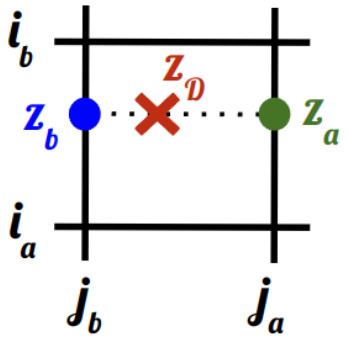


FIGURE IV.4. – Sélection des faces visibles selon un point de vue. La profondeur z_D associée aux coordonnées 2D non-entière du centre de la face F : $c_F = (x_{c_F}, y_{c_F})$.

Ensuite, nous identifions quels sommets sont visibles. Étant donné un point de vue *pov*, nous considérons qu'un sommet fait face à la caméra si au moins l'une de ses faces adjacentes est visible. Cela nous permet de réaliser un filtrage initial et d'accélérer le processus de recherche.

7.3.3. Pondération de la saillance

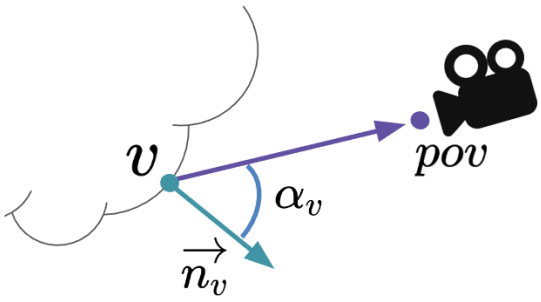


FIGURE IV.5. – Sélection des sommets visibles selon un point de vue. Cette illustration permet de visualiser de l'angle α_v entre la normale sortante du sommet v et le vecteur caméra.

Pour la fonction d'angle f, introduite dans le terme correspondant à la **saillance des sommets visibles** $\mathbf{S_a}(pov)$, nous avons testé cinq expressions différentes en fonction de l'angle α_v entre les sommets et la caméra :

- Pour favoriser les sommets faisant face à la caméra (avec $\alpha_v = 0$), nous avons utilisé : $\cos(\alpha_v)$ et $\sqrt{\cos(\alpha_v)}$.
- Inversement, pour mettre en évidence les sommets sur la silhouette, nous avons testé : $1 - \cos(\alpha_v)$ et $1 - \sqrt{\cos(\alpha_v)}$.
- Enfin, pour prendre en compte à la fois les sommets faisant face à la caméra et ceux sur la silhouette, nous avons essayé : $0.5 + \frac{1-\sqrt{\cos(\alpha_v)}}{2}$.

Pour accélérer le processus, nous calculons une seule fois le terme $\cos(\alpha_v)$, qui est présent dans toutes les versions de la formule de f, pour tous les sommets. Cette valeur de cosinus peut être calculée à partir d'un simple produit scalaire. En effet, considérons un sommet visible v et une caméra pov, comme illustré dans la Figure IV.5. L'angle α_v est celui entre la normale $\overrightarrow{n_v}$ sortante et le vecteur caméra $\overrightarrow{pov - v}$. Or, le sommet est vu depuis la position de la caméra, et de ce fait, nous sommes placés dans le repère caméra. Par définition, la caméra est au centre du repère, donc ses coordonnées sont nulles. Nous pouvons alors simplifier le vecteur $\overrightarrow{pov - v}$ en $\overrightarrow{-v}$. De plus, les deux vecteurs sont normalisés avant d'être utilisés. Ceci nous permet enfin d'écrire :

$$\cos(\alpha_v) = < \frac{\overrightarrow{n_v}}{||\overrightarrow{n_v}||} \cdot \frac{\overrightarrow{-v}}{||\overrightarrow{-v}||} >$$

En conclusion, nous avons vingt-cinq variantes différentes de l'approche proposée car nous utilisons cinq méthodes de détection de saillance 3D et cinq fonctions d'angles dif-

férentes. Cependant, une étape importante reste à mentionner : la détection des sommets et faces visibles selon un point de vue.

7.4. Proposition de validation

7.4.1. Classification des techniques de validation

Une classification des méthodes pour évaluer les performances d'une approche de type BVS est présentée dans la Table IV.1. Mais il est important de souligner que beaucoup de travaux proposent une évaluation visuels accompagnés d'arguments intuitifs, comme dans [Vázquez 01, Lee 05, Nouri 15]. Cependant, dans les approches introduites dans [Kim 17, Yang 19], des comparaisons quantitatives sont réalisées en utilisant diverses métriques, telles que le temps de calculs. Enfin, beaucoup de travaux s'appuient sur une étude utilisateurs et utilisatrices pour corréler leurs résultats avec l'opinion humaine et démontrer leur efficacité par rapport aux méthodes existantes [Leifman 16, Bonaventura 18].

Afin d'être le plus rigoureuses possible, nous avons combiné les trois aspects : évaluation qualitative, quantitative et étude utilisateurs et utilisatrices. Le § 7.4.3 ainsi que le § 8 fournissent une description de cette étude.

7.4.2. Base de données

Pour notre étude, nous avons construit une base de données comprenant 44 modèles 3D variés. Ces modèles incluent des objets humains, des créatures, des animaux, des objets du quotidien, des pièces mécaniques et des objets inconnus, cf. la Figure IV.6. Parmi ces modèles, 26 ont été fournis par la base de données associée aux travaux décrits dans [Lavoué 18]. Dans ces travaux, les auteurs ont réalisé une expérience de suivi du regard pour cartographier les zones d'objet 3D où l'attention visuelle humaine est la plus importante et d'explorer l'impact de divers facteurs sur l'attention humaine. Dans leur validation, ces auteurs se sont comparés à diverses méthodes de saillances 3D. De ce fait, ces 26 modèles 3D sont accompagnés de valeurs de saillance, générées par quatre méthodes distinctes : [Lee 05], [Song 14], [Tasse 15] et [Leifman 16]. Ensuite, les 16 autres modèles ont été acquis gratuitement en ligne et ne possédaient pas de valeurs de saillance associées. Ainsi, nous avons implémenté la méthode décrite dans [Limper 16], utilisant l'entropie de Shannon. Tout ceci nous permet ainsi d'attribuer des valeurs de saillance à l'ensemble des 44 modèles.

7.4.3. Étude utilisateurs et utilisatrices

Pour savoir si les points de vue des objets 3D, sélectionnés par notre méthode, sont bien les plus pertinents, nous avons besoin d'une vérité terrain. Cependant, dans la littérature, il n'existe pas de base de données adéquate à notre étude. Nous avons réalisé

	Comparaison avec l'état de l'art			
Travaux	Aucune	Qualitative	Quantitative	Étude utilisateur
[Blanz 99]				✓
[Vázquez 01]	✓			
[Su 21]	✓			
[Stoev 02]	✓			
[Plemenos 04]		✓		
[Sbert 05]	✓			
[Lee 05]		✓		
[Vieira 09]		✓		✓
[Vázquez 09]		✓		
[Feixas 09]	✓			
[Laga 10]	✓			
[Dutagaci 10]		✓	✓	✓
[Secord 11]				✓
[Rudoy 12]	✓			
[Habibi 15]	✓			
[Nouri 15]		✓		
[Leifman 16]		✓		✓
[Kim 17]		✓	✓	✓
[Bonaventura 18]		✓	✓	✓
[Yang 19]		✓	✓	
[Zhang 20]		✓	✓	
[Kwon 20]			✓	✓
[Schelling 21]			✓	
[Dufek 21]			✓	✓
[Marsaglia 21]			✓	✓
[Hartwig 22]		✓	✓	✓
[Biswas 23]			✓	✓
[Castelein 23]			✓	✓

TABLE IV.1. – Techniques utilisées pour la validation des méthodes de sélection de la meilleure vue d'objet 3D.

notre propre étude utilisateur à l'aide de la plate-forme Prolific [2] qui propose un service de

2. https://www.prolific.co/

FIGURE IV.6. – Exemple de modèles 3D de la base de maillages 3D utilisée.

crowdsourcing. Grâce à cette plate-forme, dédiée à la recherche, nous avons pu récolter les préférences de plus de 200 utilisateurs et utilisatrices à travers le monde. L'étude consistait à demander à chaque personne de choisir et d'ordonner 3 vues parmi les 26 proposées, pour un objet 3D, et qu'il considère comme étant les plus pertinentes et représentatives de l'objet. Chaque utilisateur a réalisé cette tâche pour dix objets 3D différents. Un exemple des répartitions des vues proposées est illustrée dans la Figure IV.7.

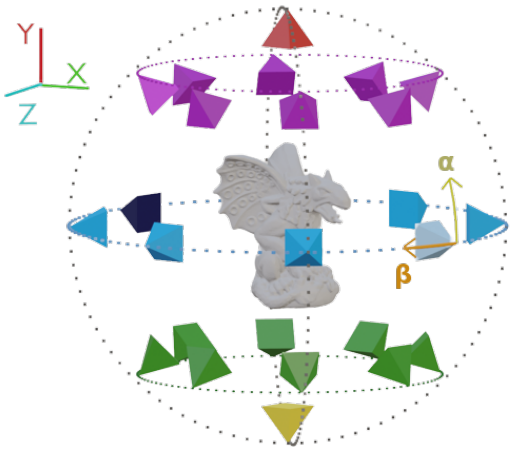


FIGURE IV.7. – Répartition des 26 points de vues sur la sphère centrée sur l'objet 3D, proposés aux utilisateurs et utilisatrices.

À l'issue de cette étude, nous avons réalisé un post-traitement qui, à partir des choix liés aux préférences humaines, construit un histogramme par objet 3D et qui indique la popularité de chacune des 26 vues. Un exemple d'histogramme lié au modèle 3D d'un sac

à dos est proposé dans la Figure IV.8.

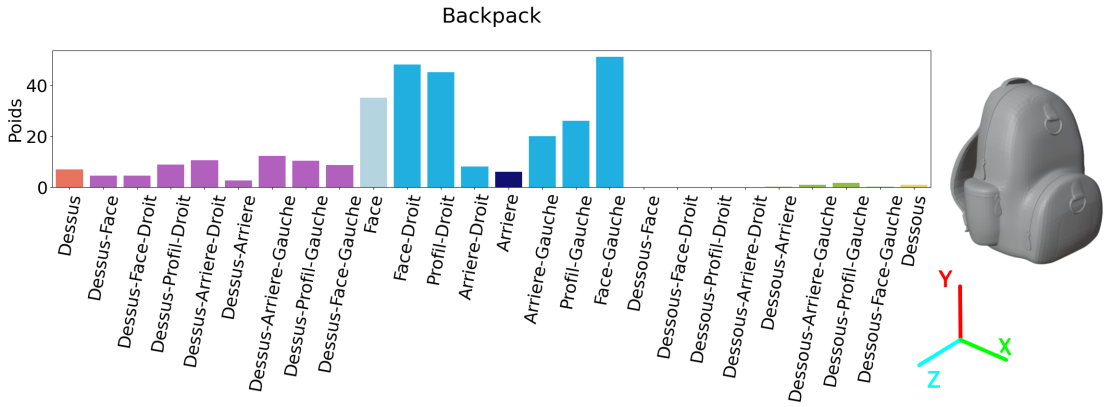


FIGURE IV.8. – Histogramme des poids de chacun des 26 points de vue étudiés, représentants leur popularité chez les utilisateurs et utilisatrices, sur le modèle 3D du sac à dos.

Tous les détails de conception de la page web pour notre étude utilisateur, le prétraitement des objets 3D, le placement des caméras, la gestion des participants, le posttraitement des données récoltées et les résultats observés sont détaillés dans le § 8.

7.4.4. Score de proximité

Pour évaluer la cohérence par rapport aux points de vue préférés par les êtres humains, nous avons défini une métrique pour déterminer un score de proximité *PS*. Ce score représente la proximité entre un point de vue donné et l'avis des utilisateurs et utilisatrices. Dans leurs travaux, les auteurs et autrices de [Dutagaci 10] ont également proposé une formule pour mesurer la distance entre deux points de vue situés sur une sphère unitaire centrée sur l'objet 3D étudié. Leur formule repose sur la distance géodésique entre les deux points de vue : celui issu d'une méthode étudiée et celui choisi par une personne. Pour calculer leur « erreur de sélection des vues », ils réalisent la moyenne des distances géodésiques obtenues pour chaque personne. Ainsi, une valeur proche de zéro signifie que l'algorithme a donné une vue proche de celle choisie par les utilisateurs et utilisatrices. Cette formule nous a inspiré le terme **C** utilisant la corrélation entre deux points de vue. Cependant, nous avons intégré un deuxième terme, nommé **W**, pour être plus robuste aux diverses situations possibles. En effet, un point de vue peut être pertinent mais pas forcément être le premier choix des utilisatrices et utilisateurs.
Plus précisément, étant donné un modèle 3D noté M, le meilleur point de vue déterminé par une méthode géométrique X est noté : $\widetilde{pov}_X^M$ et le point de vue le plus populaire chez les êtres humains est nommé : $\widetilde{pov}_U^M$. Le score de proximité entre ces deux points de vue est définie par :

$$PS_X^M = \max(\mathbf{C}(\widetilde{pov}_X^M, \widetilde{pov}_U^M), \mathbf{W}(\widetilde{pov}_X^M)) \tag{IV.5}$$

Avec :

- $\mathbf{C}(\widetilde{pov}_X^M, \widetilde{pov}_U^M)$ la distribution normale en 3D [Chave 15], centrée en $\widetilde{pov}_U^M$, avec un support $\sigma = 1.3$ et translatée sur l'intervalle $[0, 1]$. Ce terme mesure la quantité d'informations communes entre les deux points de vue donnés en entrée. Autrement dit, ce terme quantifie les informations qu'ils partagent. L'impact de la valeur numérique choisie pour σ est illustré dans la Figure IV.9.

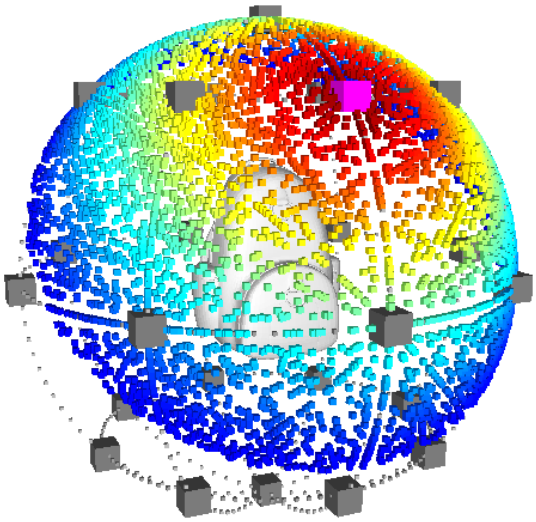


FIGURE IV.9. – Zone d'influence du score de proximité. La représentation de la distribution normale 3D centrée sur la caméra magenta et associée au terme $\mathbf{C}$ de la formule du score de proximité IV.5.

- $\mathbf{W}(\widetilde{pov}_X^M)$ le poids associé à $\widetilde{pov}_X^M$ dans l'histogramme du modèle M, divisé par le poids maximal de l'histogramme pour être compris dans l'intervalle $[0, 1]$.

Dans le cas où $\widetilde{pov}_X^M$ ne correspond pas exactement à $\widetilde{pov}_U^M$, il existe deux situations où $\widetilde{pov}_X^M$ est considéré comme pertinent pour l'opinion humaine. Premièrement, il est possible que $\widetilde{pov}_X^M$ soit un point de vue peu populaire auprès des utilisateurs et utilisatrices, autrement dit, qu'il ait un poids faible dans l'histogramme des popularités du modèle M mais ce point de vue peut être sémantiquement très proche de $\widetilde{pov}_U^M$: deux points de vue juxtaposés sur la sphère peuvent partager un grand nombre de caractéristiques visibles. Ainsi, un point de vue $\widetilde{pov}_X^M$ contenu dans le voisinage proche du point de vue des utilisateurs et utilisatrices $\widetilde{pov}_U^M$ sera considéré comme acceptable proportionnellement à sa distance par rapport à la position exacte de $\widetilde{pov}_U^M$. C'est au travers du terme $\mathbf{C}(\widetilde{pov}_X^M, \widetilde{pov}_U^M)$ que nous parvenons à considérer cette situation comme acceptable, comme illustré dans la Figure IV.10a. Deuxièmement, $\widetilde{pov}_X^M$ peut-être un point de

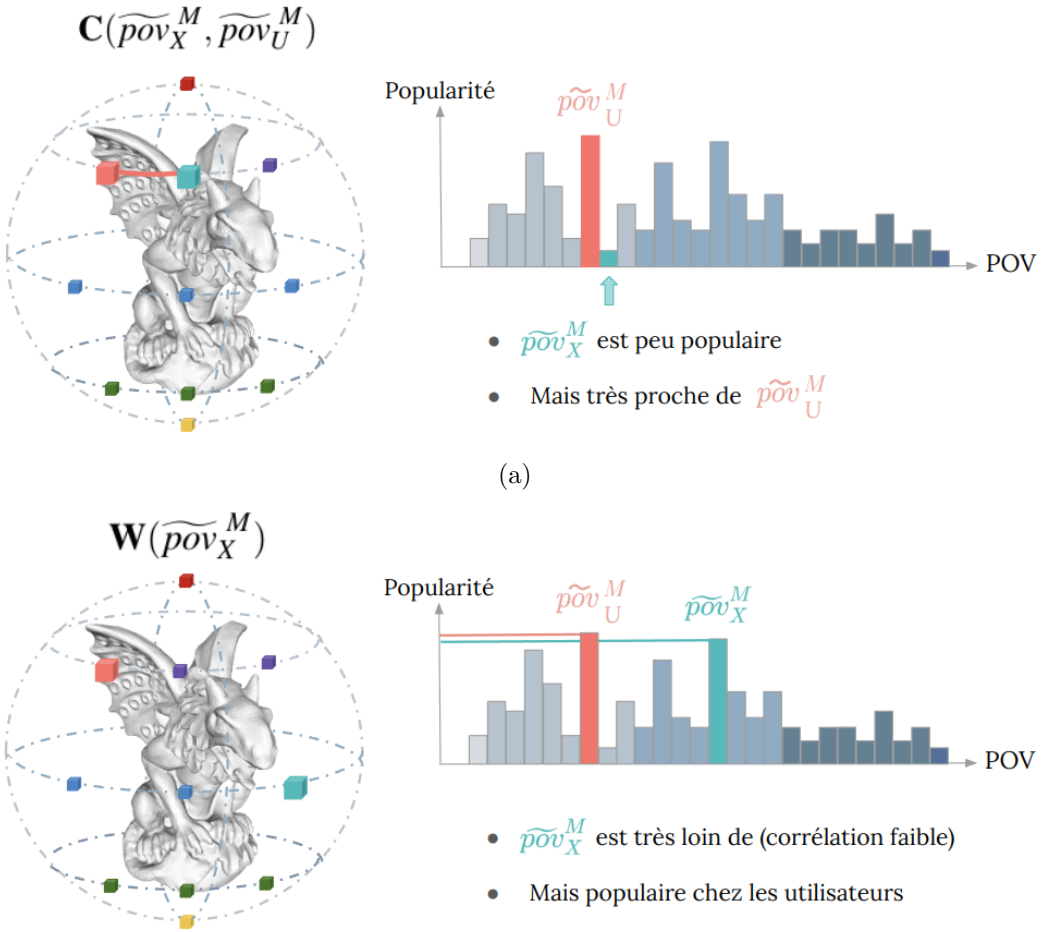


FIGURE IV.10. – Illustration des vues particulières considérées comme pertinentes. Nous présentons deux situations avec lesquelles la vue déterminée par une méthode géométrique est pertinente alors qu'elle ne correspond pas à celle choisie par les utilisateurs et utilisatrices.

vue presque aussi populaire que $\widetilde{pov}_U^M$, ce qui correspond à un point de vue significatif. Cette situation est illustrée dans la Figure IV.10b. Après plusieurs tests empiriques avec différentes combinaisons, l'utilisation du maximum nous permet de considérer ces deux situations.

Cette métrique d'évaluation est utilisée dans le reste de ce chapitre, à la fois pour choisir la meilleure configuration de la méthode proposée et pour comparer les approches de l'état de l'art.

7.5. Optimisation de notre approche

7.5.1. Choix de la méthode de saillance 3D

Afin de fixer notre terme de saillance des sommets visibles $\mathbf{S_a}(pov)$, nous devons choisir entre les différentes méthodes de saillance 3D pour le terme S_i et les fonctions d'angle f, discutées dans la section 7.3, Pour ce faire, nous utilisons le score de proximité, précédemment défini. En effet, nous calculons le score de proximité entre les meilleures vues obtenues à partir de chacune des vingt-cinq versions disponibles de la formule proposée IV.4 et les vues préférées des utilisateurs et utilisatrices, pour tous les modèles de notre étude.
Nous rappelons que nous utilisons les modèles 3D de la base de données fournies dans [Lavoué 18] avec des valeurs de saillances issues de quatre méthodes de l'état de l'art : [Lee 05], [Song 14], [Tasse 15], [Leifman 16].

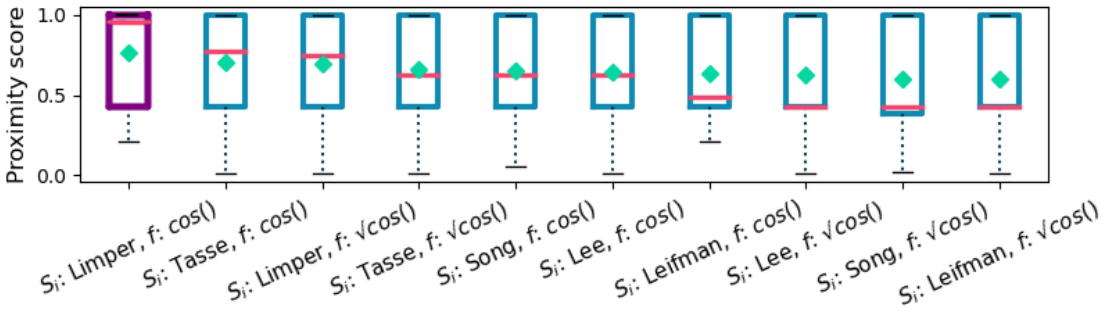


FIGURE IV.11. – Résultats quantitatifs de l'étape d'optimisation. Nous présentons les données statistiques des scores de proximité *PS* obtenus par les dix meilleures versions de la formule proposée IV.4, en spécifiant la méthode de saillance intrinsèque S_i et la fonction d'angle f utilisées dans le terme $\mathbf{S_a}$. Les scores de proximité sont calculés sur les 26 modèles de [Lavoué 18].

La Figure IV.11 affiche la répartition des scores de proximité pour les dix meilleures combinaisons testées pour le terme proposé $\mathbf{S_a}(pov)$. Grâce à l'utilisation des diagrammes en boîtes, il est facile d'observer que la combinaison la plus performante est celle utilisant [Limper 16] [3] pour calculer le terme S_i et la fonction $\cos(\alpha_v)$ pour la fonction d'angle f. En effet, avec cette combinaison, 25% des vues sélectionnées par notre formule correspondent exactement à celles fournies par l'étude utilisateurs et utilisatrices. Ce résultat est cohérent avec les conclusions émises par les travaux réalisant des bilans de toutes les mesures de sélection de la meilleure vue possible. En effet, les attributs utilisant l'entropie ainsi que la courbure, comme ceux utilisés dans [Limper 16], font partie des meilleurs dans les études comparatives réalisées dans [Dutagaci 10, Secord 11, Bonaventura 18].

3. [Limper 16] a proposé une méthode d'estimation de la saillance en 3D et non de la meilleure vue.

Nous pouvons remarquer que les fonctions d'angle f dans le top 10 sont celles qui mettent en évidence les sommets faisant face à la caméra. Nous pouvions nous attendre à ce que la silhouette donne des informations importantes, mais le fait de privilégier la silhouette favorise probablement aussi les vues accidentelles : celles avec un problème de perspective qui réduit la compréhension globale de l'objet, comme dans la Figure IV.1b.

7.5.2. Étude d'ablation

Maintenant que la version la plus optimale du terme $\mathbf{S_a}(pov)$ a été déterminée, il est nécessaire de vérifier que chacun des trois termes contribue de manière positive et significative dans la recherche de points de vue cohérents avec la logique humaine. Pour montrer l'importance de chaque terme de notre formule IV.4, nous réalisons une étude d'ablation ou *ablation study*. La contribution de chaque terme est présentée dans la Figure IV.12. Le diagramme de gauche, correspondant à notre formule IV.4 originale ($\mathbf{S} + \mathbf{S_a} + \mathbf{S_e}$), a une moyenne de $0,765$, tandis que la seconde ($\mathbf{S_a} + \mathbf{S_e}$) a une moyenne de $0,746$. Ainsi, la visibilité de surface $\mathbf{S}$ contribue, mais les termes $\mathbf{S_a}$ et $\mathbf{S_e}$ ont un impact plus important. Même si le terme $\mathbf{S_a}$ est plus important en moyenne, il faut noter que seuls les modèles avec un visage (et donc possédant des yeux) voient leur score de proximité modifié par le terme $\mathbf{S_e}$. Avec l'ajout du terme $\mathbf{S_e}$, la médiane passe de 0.627 à 0.955. En conclusion, cette étude d'ablation confirme l'intérêt de chacun des trois termes proposés et ils seront donc conservés.

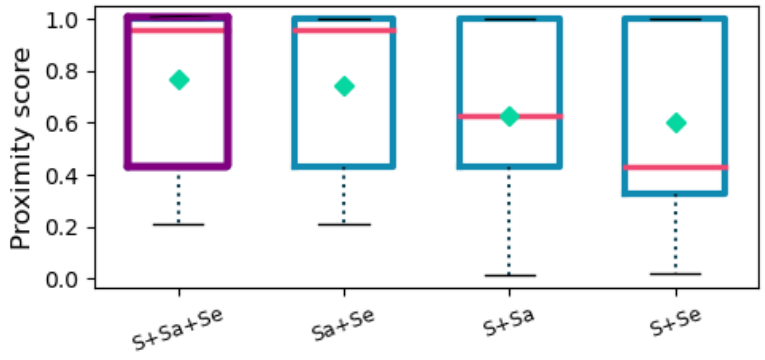


FIGURE IV.12. – Étude d'ablation. Elle montre que les trois termes $\mathbf{S}$, $\mathbf{S_e}$ et $\mathbf{S_a}$ contribuent à la qualité du résultat. En particulier, $\mathbf{S}$ améliore la moyenne de $\sim 0,02$.

7.6. Comparaisons quantitatives et qualitatives

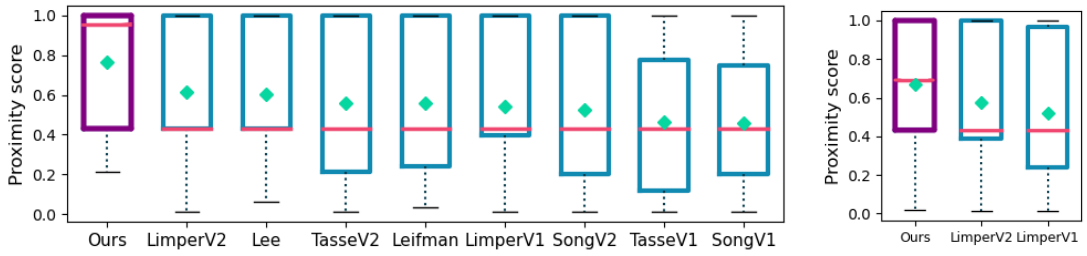
Pour évaluer l'efficacité de notre formule IV.4, nous avons considéré deux méthodes purement géométriques de l'état de l'art : [Lee 05] et [Leifman 16]. Pour rappel, voici les formules de sélection du meilleur point de vue pour ces deux méthodes : soit un modèle

3D M et un point de vue *pov* avec son ensemble V de sommets visibles, le score attribué à ce point de vue est défini par :

- [Lee 05] : $\sum_{v \in V} \text{Saillance}(v)$
- [Leifman 16] : $\sum_{v \in V} \text{Saillance}(v) * \sqrt{\cos(\alpha_v)}$

Chacun de ces travaux introduit sa propre méthode de calculs de saillance 3D.

Comme mentionné précédemment, la base de données que nous utilisons nous fournit les valeurs de saillance des sommets pour 26 modèles issus de deux autres méthodes : [Tasse 15] et [Song 14]. Pour réaliser des comparaisons plus complètes, nous avons décidé d'utiliser ces valeurs de saillance en les injectant dans les deux formules de sélection de la meilleure vue proposées dans [Lee 05] et [Leifman 16]. Ces nouvelles combinaisons forment deux nouvelles méthodes appelées respectivement : TasseV1 et SongV1 (TasseV2 et SongV2) lorsque la formule de sélection des points de vue utilisée est celle de [Lee 05], suivant la formule de [Leifman 16]. Une fois la saillance calculée pour chaque sommet des modèles, nous utilisons également les formules de sélection de la meilleure vue introduites dans [Lee 05] et [Leifman 16] pour obtenir deux nouvelles méthodes géométriques, respectivement nommée LimperV1 et LimperV2.



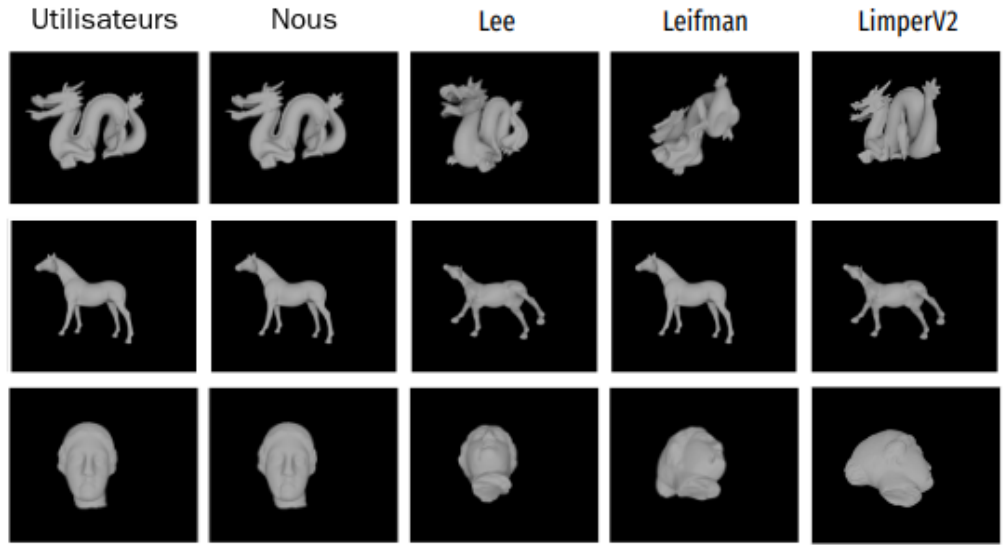
(a) (b)

FIGURE IV.13. – Comparaisons quantitatives des méthodes étudiées. Nous présentons les données statistiques des scores de proximité *PS* du meilleur (à gauche) au plus mauvais (à droite) : en (a), calculés sur 25 modèles proposés dans [Lavoué 18] et en (b), l'ensemble des 43 modèles (le modèle *prot* a été retiré).

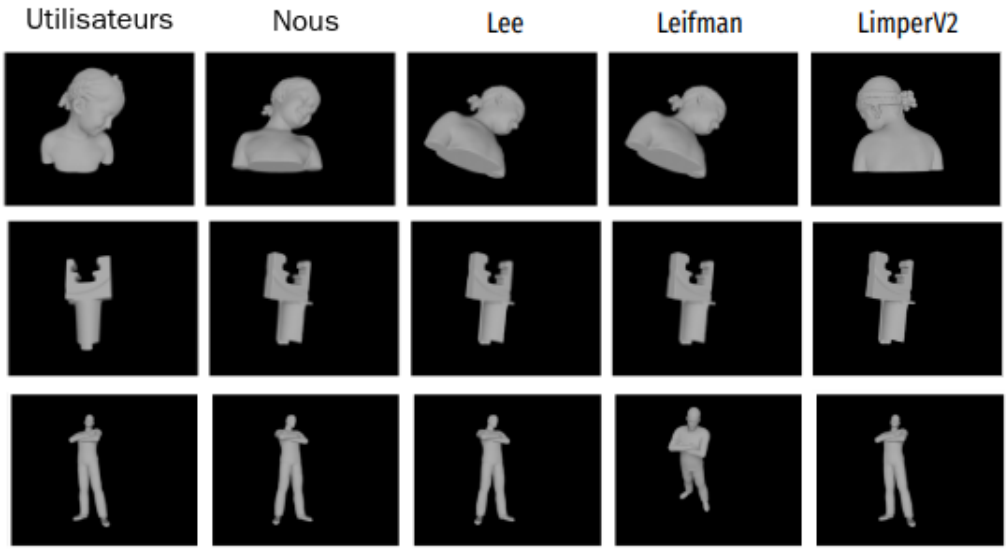
Chaque modèle possède un meilleur point de vue pour chacune des neuf méthodes étudiées (notre méthode et les huit auxquelles nous souhaitons nous comparer). Les scores de proximité *PS* de chaque modèle pour chaque méthode sont présentés dans la Figure IV.13. Dans un premier temps, après avoir étudié les résultats obtenus par le modèle *prot* (en haut à droite de la Figure IV.6), aucune vue n'était plus pertinente qu'une autre : ce modèle a donc été retiré de notre étude. En effet, ce modèle 3D correspond à une molécule et n'a donc, par définition, aucune vue plus représentative qu'une autre pour les personnes qui ne sont pas expertes.

Dans la Figure IV.13a, nous affichons les résultats des scores de proximité pour la meilleure version de notre formule et les huit autres méthodes, calculés sur les modèles proposés dans [Lavoué 18]. Comme nous pouvons l'observer, le diagramme de notre méthode est le meilleur : 50% des modèles obtiennent un score supérieur à 0.955, alors que les autres méthodes ont une médiane de 0.43 au maximum. En moyenne, notre méthode obtient un score de proximité de 0.76, alors que les autres ne dépassent pas 0.61. Les mêmes résultats sont obtenus en considérant l'ensemble des modèles étudiés dans la Figure IV.13b. Celle-ci présente les résultats statistiques pour notre méthode et les deux autres méthodes géométriques applicables sur l'ensemble de nos 44 modèles : LimperV1 et LimperV2. Même avec un plus large nombre de modèles 3D, notre méthode reste la plus pertinente, avec 50% des scores de proximité supérieurs à 0.7.

Par ailleurs, quelques résultats visuels sont présentés dans la Figure IV.14. Nous pouvons voir que certains de nos points de vue, illustrés dans la deuxième colonne de la Figure IV.14a, sont exactement les mêmes que ceux issus de l'étude utilisatrice et utilisateur, illustrés dans la première colonne de la Figure IV.14a, contrairement aux trois autres méthodes géométriques. Cependant, dans certains cas, notre méthode ne parvient pas à retrouver exactement le même point de vue que celui préféré par les êtres humains. Trois exemples sont mis en avant dans la Figure IV.14b. Par exemple, dans la première ligne avec le modèle *Bimba*, notre point de vue peut être considéré comme moins esthétique, mais c'est intentionnel, car il met en valeur les traits du visage, en particulier les yeux. Ce résultat est donc cohérent avec nos attentes. Deuxièmement, le point de vue choisi par les personnes pour la pièce mécanique de la deuxième ligne est accidentel. En effet, avec cette vue, il y a une ambiguïté sur la profondeur de l'objet. Cette ambiguïté de perspective est éliminée dans les vues issues des quatre méthodes géométriques. Enfin, le point de vue choisi par les utilisatrices et utilisateurs pour l'homme aux bras croisés est également esthétique et propre à la culture humaine alors que notre point de vue met mieux en valeur les détails de la position : les bras croisés et le visage. Aucune des méthodes ne parvient à identifier exactement le même point de vue que celui préféré par les utilisatrices et utilisateurs. Cependant, les points de vue déterminés peuvent être considérés comme symétriques à celui attendu, même si l'objet lui-même n'est pas parfaitement symétrique.



(a)



(b)

FIGURE IV.14. – Comparaisons qualitatives des méthodes étudiées. Il s'agit d'un sous-ensemble des meilleurs points de vue sélectionnés par les personnes (première colonne de (a) et (b)), puis par la méthode proposée (deuxième colonne) et par trois autres méthodes géométriques (de la troisième à la cinquième colonne). Avec certains modèles, nous avons sélectionné la même vue que les utilisateurs et utilisatrices (a), alors que sur d'autres les points de vue sont différents (b) mais les résultats restent corrélés. Bien souvent, le point de vue que nous avons sélectionné est plus informatif.

En résumé

Cette section se concentre sur la sélection du meilleur point de vue d'un objet à partir de son maillage 3D. Tout d'abord, nous avons montré qu'étant donné le contexte, une approche basée apprentissage n'est pas complètement adaptée. Ainsi, nous proposons une méthode purement géométrique pour déterminer le point de vue le plus pertinent et représentatif d'un objet, sans se préoccuper de l'esthétisme. La méthode s'appuie sur trois termes évaluant les quantités de surface visibles et la saillance intrinsèque. Ce dernier utilise une pondération spécifique pour chaque sommet visible, dépendant de l'angle de vue, permettant de favoriser les points les plus saillants qui constituent l'information essentielle de l'objet étudié. Ces trois termes sont sommés pour former un score attribué à chaque point de vue, le plus pertinent étant celui avec le score le plus élevé. Diverses combinaisons ont été testées, en explorant différentes méthodes de calculs de saillance intrinsèque 3D et des pondérations pour obtenir la version optimale de l'approche.
Pour valider ces résultats, les préférences humaines ont été recueillies via une étude utilisatrice et utilisateur détaillée dans la section suivante. Les vues considérées comme les plus pertinentes doivent être cohérentes avec la logique et la perception humaine, et offrir une représentation significative.
En outre, pour comparer les résultats, deux méthodes de l'état de l'art ont été étudiées et six autres méthodes ont été élaborées à partir des données disponibles. Les résultats quantitatifs et qualitatifs démontrent l'efficacité et la pertinence de la méthode proposée par rapport aux autres approches évaluées.

Ces travaux ont été présentés lors de la JFIG (*Journées Françaises de l'Informatique Graphique*), en 2023, puis publiés récemment lors de la conférence internationnale VISAPP (*Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*) [Pelissier-Combescure 24].

8. Réalisation d'une étude utilisateurs et utilisatrices

8.1. Contexte

Lorsqu'il s'agit de mettre en valeur un objet 3D, comme un humain, il est essentiel de comprendre les préférences de ces derniers. Comme mentionnée dans le paragraphe 7.4.1 et illustré dans la Table IV.1, certaines méthodes de sélection de la meilleure vue d'un objet ont mené des études utilisateurs pour vérifier la cohérence entre leurs résultats et les préférences humaines. Pour élaborer notre propre étude utilisateur, nous avons étudié en détails ces études, que nous avons résumées dans la Table IV.2. Nous avons uniquement examiné les études similaires à la nôtre en termes de données. Par exemple, l'étude décrite dans [Castelein 23] aborde la sélection du meilleur point de vue pour les projections 3D de nuage de points, après une réduction de dimension des données.

Deux catégories d'études se distinguent : celles qui laissent la possibilité aux utilisateurs de naviguer librement autour des objets avant d'émettre leur(s) choix [Blanz 99, Dutagaci 10], et celles qui proposent un nombre limité d'images pour faire la sélection d'une ou de plusieurs vues [Secord 11, Leifman 16, Kim 17, Hartwig 22]. Nous avons décidé de proposer un intermédiaire : les utilisateurs peuvent naviguer autour de l'objet et l'appréhender dans son intégralité avant de faire leurs choix, mais sont limités par un nombre fixé de points de vue. offre ainsi aux utilisateurs la possibilité de choisir un point de vue approprié.

Demander aux utilisateurs une seule vue ne permet pas de réaliser une analyse et une comparaison réellement rigoureuse et interprétable. Pour avoir une meilleure précision et une distribution de la popularité des points de vue sélectionnés, plus fines et représentatives de la réalité, nous nous sommes inspirées des travaux détaillés dans [Leifman 16, Kim 17, Hartwig 22]. En effet, nous avons demandé à nos utilisateurs de sélectionner non pas un mais trois points de vue qu'ils trouvent pertinents pour identifier les objets. Comme illustré dans la section 7.4, grâce à cette multiplicité des données récoltées, nous avons pu réaliser une étape de validation plus précise.

Travaux	Données	Technique	Participants	Instructions
[Blanz 99]	14 objets 3D	Navigation en continue	36	« Quelle vue choisiriez-vous pour donner la meilleure impression d'un objet dans une brochure ? »
[Dutagaci 10]	68 objets 3D	Navigation en continue	26	« Sélectionnez la vue la plus informative de l'objet. »
[Secord 11]	16 objets 3D avec chacun 120 paires d'images	Comparaison par paire	524	« Laquelle de ces deux vues préférez-vous ? »
[Leifman 16]	79 objets 3D avec chacun 12 images	Sélection d'images	195	« Indiquez les vues qui permettent de comprendre la forme de l'objet. »
[Kim 17]	100 objets 3D avec chacun 8 images	Sélection d'images	52	"Sélectionnez la meilleure vue, puis la deuxième meilleure vue. »
[Kwon 20]	25 objets 3D (humains)	Navigation en continue	23	« Identifiez celle que vous jugez préférable. »
[Hartwig 22]	3220 objets 3D avec chacun 6 triplets d'images	Sélection d'une image	950	« Sélectionnez parmi trois images celle qui présente la meilleure et la pire vue parmi les huit images. »

TABLE IV.2. – Détails techniques utilisés lors des études utilisateurs des méthodes de sélection de la meilleure vue d'objet 3D.

Nous avons réalisé notre étude utilisateur à l'aide de la plateforme *Prolific*[4], qui offre des services de *crowdsourcing*. Notre souhait est de concevoir une interface permettant de recueillir les avis des utilisateurs sur une série de modèles 3D, en identifiant les points de vue jugés les plus représentatifs selon eux. Cette étude a été encadrée par un protocole expérimental rigoureux conçu pour assurer des résultats fiables et significatifs.

Objectif

Récolter les trois vues considérées comme étant les plus pertinentes pour un objet 3D, selon les utilisateurs et utilisatrices de l'étude.

Ce protocole a été élaboré pour guider les participants tout au long de l'expérience, en leur permettant d'évaluer les différents modèles 3D et de fournir leurs commentaires. Les différentes étapes de ce protocole sont décrites dans la section 8.2. Plus précisément, nous commençons par présenter de manière détaillée l'étude utilisateur, cf. § 8.2.1, et la définition, que nous avons fournie lors de l'étude, de ce que nous considérons comme étant un « bon » point de vue dans nos travaux cf. § 8.2.2. L'activité principale ainsi que le tutoriel proposé sont mentionnés dans la section 8.2.3 et 8.2.4. Le processus de stockage des données pour assurer une gestion appropriée des informations collectées est décrit dans la section 8.2.5.

Une attention particulière a été accordée à la conception de l'interface, intégrant diverses fonctionnalités essentielles pour une collecte précise et efficace des préférences des utilisateurs. Ces fonctionnalités concernent principalement la mise en place d'outils interactifs, tels que des boutons ou des flèches, pour permettre aux utilisateurs de réaliser le plus efficacement et agréablement possible les tâches à accomplir. Les détails de cette conception, ainsi que la description des fonctionnalités, sont exposés dans la section 8.3. Les aspects techniques, tels que le prétraitement des modèles 3D ou le choix des vues proposées aux utilisateurs, sont décrits dans la section 8.4.

Une fois que les données ont été récoltées, il est nécessaire de les traiter et d'en juger la qualité. Pour cela, nous avons procédé à un processus de post-traitement des données, détaillé dans la section 8.5. Chaque étape, de la réception des données, cf. § 8.5.1, à l'analyse des données, cf. § 8.5.2, en passant par le filtrage des participants, cf. § 8.5.3, et les calculs de la popularité, de chacune des vues proposées aux utilisateurs pour chaque modèle 3D, cf. § 8.5.4, est expliquée en détail.

Enfin, les résultats de cette étude, révélant les points de vue considérés comme étant les meilleurs par les utilisateurs, sont présentés dans la section 8.6. Cette analyse approfondie

4. https ://www.prolific.co/

nous permet de réaliser des conclusions significatives sur les tendances et les préférences des utilisateurs. Plus précisément, nous réalisons des interprétations sur les préférences humaines dans un contexte général, cf. § 8.6.1, puis en tenant compte de certains aspects spécifiques liés aux modèles 3D eux-mêmes ou bien aux vues. Par exemple, nous avons étudié la répartition des vues accidentelles, cf. § 8.6.2, ou occultées, cf. § 8.6.3, dans les choix des utilisateurs. Nous avons également analysé les choix des utilisateurs lorsque les modèles possèdent des yeux, cf. § 8.6.4, sont symétriques, cf. § 8.6.5, ou ne sont pas familiers, cf. § 8.6.6.

Enfin, nous concluons cette section par des comparaisons entre les préférences des utilisateurs lorsque que nous leur présentons des objets 3D sans texture (comme dans l'étude) et dans la vie courante, au travers d'images réelles, présentées dans la section 8.7

8.2. Présentation de l'interface

8.2.1. Page d'introduction

L'étude utilisateur, que nous avons élaborée, commence par la définition d'un point de vue d'un objet en 3D. Celle-ci a pour objectif de rappeler que le point de vue représente la position à partir de laquelle nous observons l'objet. Cette position spécifique révèle certaines parties de l'objet, tout en occultant d'autres. Pour clarifier cette idée, nous avons illustré notre définition par un exemple concret : une vue de face d'un loup, suivie d'une vue de dos du même loup. Ces deux perspectives distinctes mettent en lumière des aspects différents de l'objet. Les pages de notre interface associées à cette définition sont affichées dans la Figure IV.15.

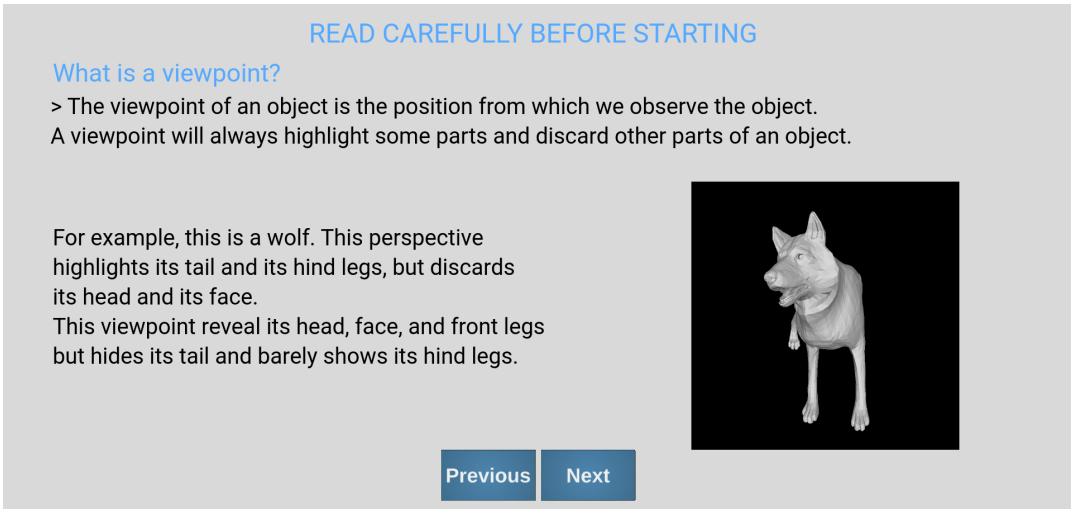
8.2.2. Définition du « bon » point de vue

Après avoir introduit cette définition, nous avons explicité ce que nous considérons comme étant un « bon » point de vue. Selon notre définition, un bon point de vue est celui qui offre une vision pertinente de l'objet. Il doit être à la fois représentatif de l'objet et mettre en valeur ses caractéristiques essentielles. Ainsi, la question principale de notre étude qui s'affiche est :

Problématique

« Quel point de vue préférez-vous pour mettre en valeur et reconnaître cet objet ? »

(a)



(b)

FIGURE IV.15. – Illustration de la définition d'un point de vue : en fonction de ce dernier, certaines parties sont soient visibles, soient cachées.

8.2.3. Tutoriel proposé et consignes

Après l'inscription des participants, un tutoriel est mis à leur disposition. Ce tutoriel vise à familiariser les utilisateurs et utilisatrices avec toutes les fonctionnalités de l'interface, notamment la manière de manipuler les objets pour les observer sous différents angles, de sélectionner des points de vue, et de modifier leurs choix. Ensuite, la partie principale de l'étude débute. Chaque participant a la tâche d'étudier 10 modèles 3D distincts et de sélectionner trois points de vue pour chaque modèle. Plus précisément, l'utilisateur doit ordonner ses trois choix, en commençant par celui qui répond le plus fidèlement à la problématique énoncée, suivi du deuxième choix, puis du troisième. Les 10 modèles sont choisis aléatoirement dans l'ensemble de données et sont initialement affichés sous un point

de vue aléatoire. L'interface correspondant à cette étape est illustrée dans la Figure IV.16.

8.2.4. Justification des choix des utilisateurs

Une fois cette phase de sélection terminée, chaque participant est invité à justifier ces choix pour un sous-ensemble de 5 des 10 modèles précédemment étudiés. Ces 5 modèles sont également sélectionnés au hasard. Pour aider les participants à justifier leurs choix, nous leur proposons des mots clés tels que "Vue latérale, Vue de face, Vue d'ensemble, Contact visuel, Agréable, Reconnaissable, Vue de 3/4". De plus, un champ de texte libre est disponible pour leur permettre d'entrer des raisons spécifiques.

8.2.5. Stockage des données

Enfin, toutes les données collectées sont soigneusement stockées sur un serveur dédié. Chaque participant génère un fichier JSON qui est ensuite mis à disposition sur notre plateforme de l'IRIT : `https://gitlab.irit.fr/bvs-study/data.git`. Ce processus garantit une gestion rigoureuse et sécurisée des données recueillies au cours de l'étude utilisateur.

Après avoir établi le protocole expérimental rigoureux qui a guidé notre étude, nous allons maintenant nous attarder sur les détails de l'interface que nous avons élaborée spécifiquement pour permettre aux participants de mettre ce protocole en action de manière fluide et efficace.

8.3. Fonctionnalités de l'interface

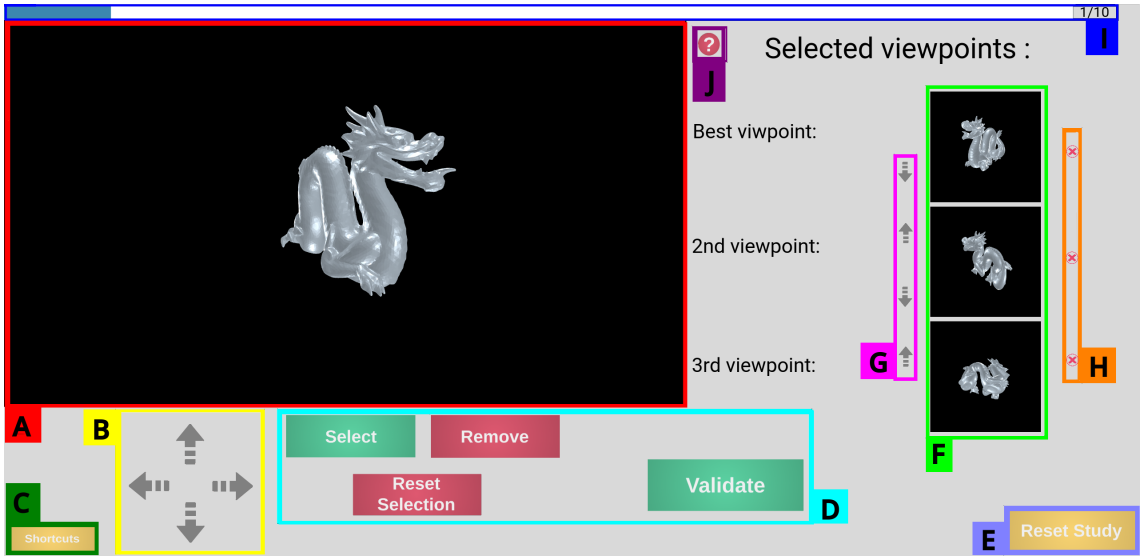


FIGURE IV.16. – Interface principale permettant aux utilisateurs de visionner les modèles 3D et de sélections leurs trois points de vue.

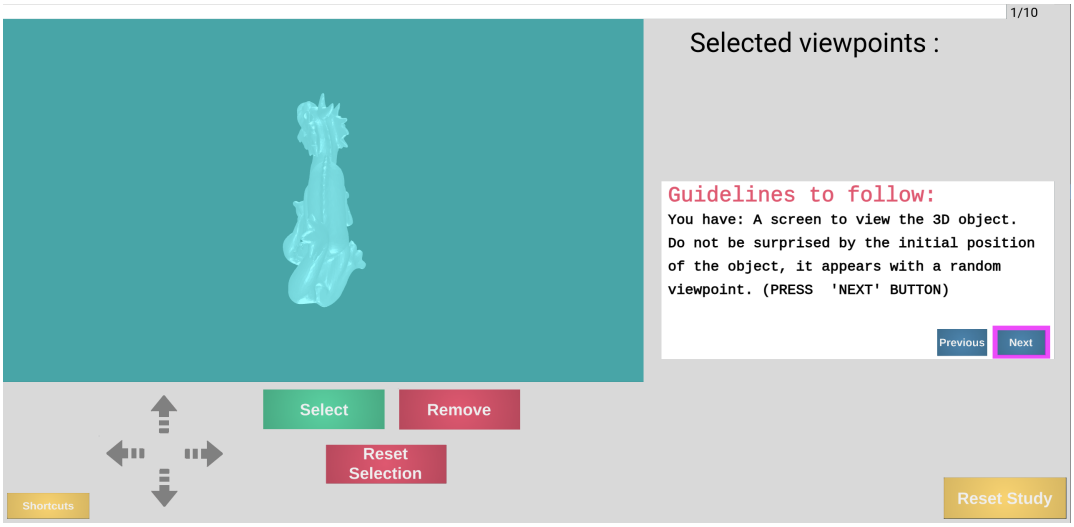
Le cœur de notre étude est la partie qui permet aux utilisateurs de visualiser les modèles 3D et de sélectionner les trois points de vue qu'ils considèrent les plus adaptés pour répondre à la problématique. Notre objectif est de rendre cette expérience aussi agréable que possible tout en garantissant sa simplicité, son efficacité et son accessibilité.

Dans cette section, nous présentons en détail les différentes fonctionnalités de cette interface, notamment les zones informatives qui guident les participants, les éléments interactifs tels que les flèches de navigation et les boutons d'action, ainsi que les boutons informatifs à survoler. La Figure IV.16 illustre visuellement l'apparence de notre interface et servira de référence pour mieux comprendre les actions et fonctionnalités que nous allons expliciter dans la Table IV.3 ci-dessous.

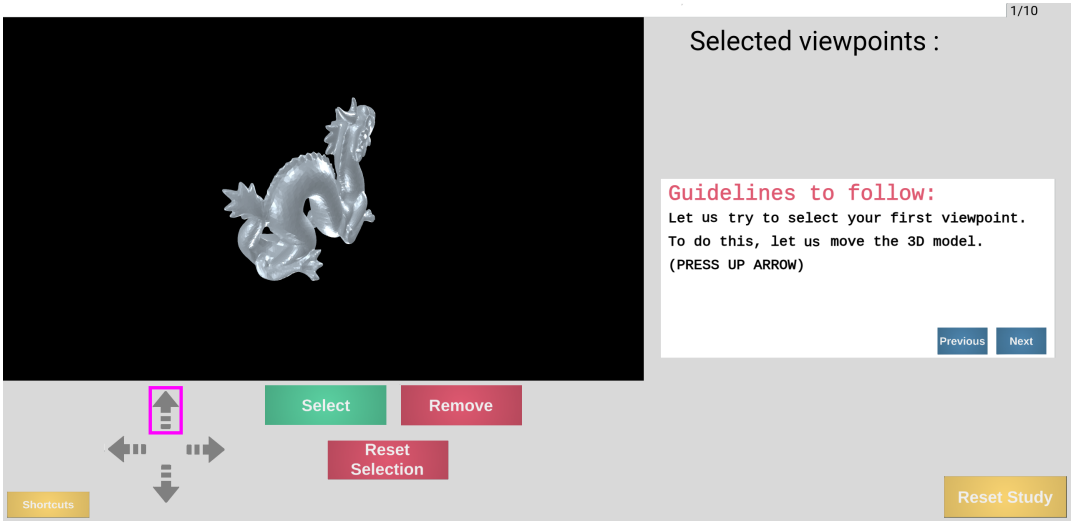
Zone	Type	Description
A	Visuel	Ecran pour visualiser le modèle 3D dans un environnement neutre.
B	Flèches cliquables	Flèches permettant de déplacer la caméra autour du modèle 3D et ainsi observer différents points de vue du modèle.
C	Bouton à survoler	Bouton qui affiche sur l'écran un clavier avec les raccourcis disponibles mis en évidence
D	Boutons cliquables	Ensemble de boutons qui permettent soit de sélectionner le point de vue courant, soit de retirer la dernière vue sélectionnée, soit de retirer toutes les vues sélectionnées, soit de valider la sélection.
E	Bouton à survoler	Bouton qui permet de reprendre la sélection des points de vue depuis le premier modèle 3D.
F	Visuel	Récapitulatif visuel des points de vue déjà sélectionnés pour le modèle 3D courant.
G	Flèches cliquables	Flèches permettant d'intervertir l'ordre des points de vue déjà sélectionnés pour le modèle 3D courant.
H	Boutons cliquables	Boutons en croix qui permettent de supprimer un point de vue déjà sélectionné pour le modèle 3D courant tout en conservant l'ordre des vues restantes.
I	Visuel	Bar de progression informant du nombre de modèles 3D restant à étudier au cours de l'étude.
J	Bouton à survoler	Bouton d'aide qui affiche un récapitulatif des consignes lorsque l'utilisateur le survole.

TABLE IV.3. – Description de chacune des fonctionnalités disponibles sur la page principale de l'interface.

Chacune des fonctionnalités de l'interface est expliquée au moyen d'un tutoriel intégré, conçu pour faciliter la compréhension des utilisateurs à travers des interactions ludiques et guidées. En effet, le tutoriel offre des descriptions détaillées des boutons et des flèches annexes, fournissant aux utilisateurs toutes les informations nécessaires pour une utilisation optimale. Par exemple, dans la Figure IV.17a, la page du tutoriel met en avant l'écran noir qui affiche le point de vue courant de la caméra. Certaines pages du tutoriel ont été conçues pour faire intervenir les utilisateurs dans leur apprentissage de l'outil. En effet, les utilisateurs sont obligés d'interagir avec l'interface en cliquant sur les flèches et les boutons indiqués pour pouvoir poursuivre le tutoriel.



(a) Page du tutoriel : mis en évidence de l'écran pour observer les modèles 3D.



(b) Page du tutoriel : manipulation de l'objet 3D à l'aide des flèches directionnelles.

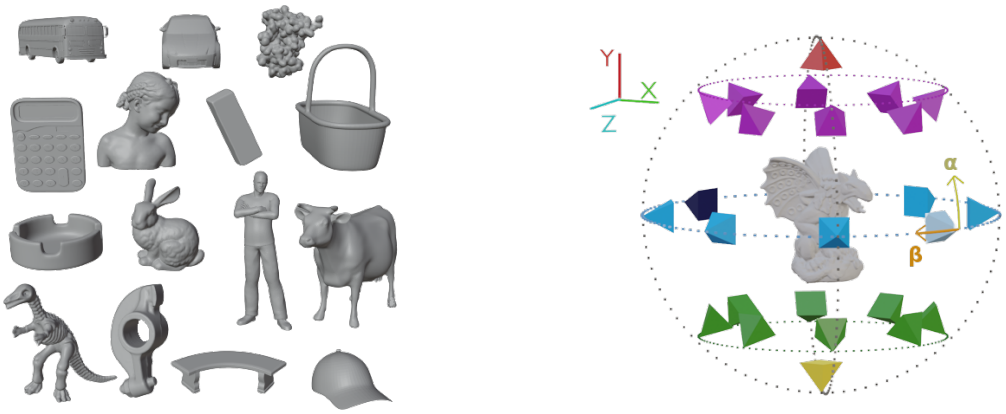
FIGURE IV.17. – Exemple de deux pages du tutoriel proposé dans l'étude utilisateur.

La Figure IV.17b affiche une page du tutoriel qui oblige l'utilisateur à cliquer sur la

flèche encadrée en magenta pour changer la position de la caméra et ainsi afficher un nouveau point de vue du modèle 3D courant. Ces actions offrent ainsi une expérience d'apprentissage interactive et pratique. Maintenant que nous avons énuméré les différentes fonctionnalités de notre interface, nous allons détailler les aspects techniques du pré-traitement des modèles 3D que les utilisateurs visualisent, explorant ainsi les étapes cruciales qui garantissent une expérience visuelle optimale.

8.4. Pré-traitement des caméras et des modèles 3D

Notre interface est développée en utilisant le langage de programmation *JavaScript* et utilise la librairie *Three.js* pour le traitement et l'affichage des modèles 3D. Cette librairie offre une gamme d'outils efficaces pour gérer la visualisation 3D, garantissant ainsi une expérience utilisateur fluide et réaliste.



(a) (b)

FIGURE IV.18. – (a) Visualisation d'un échantillon des modèles 3D de notre base de données, (b) Répartition des caméras d'étude sur la sphère centrée sur le centre du modèle 3D. Le code couleur permet de visualiser les différents ensembles de caméras.

Notre base de données de modèles 3D se compose de 44 modèles réguliers. Parmi ces modèles, 26 ont été acquis auprès de [Lavoué 18], tandis que les 18 autres ont été récoltés depuis internet. Certains de ces modèles sont illustrés dans la Figure IV.18a. Chacun de ces modèles a été préparé pour garantir une cohérence dans l'étude. Ils ont été positionnés de manière à avoir l'axe Y orienté de bas en haut, suivant le sens de la gravité lorsque cela est pertinent. De plus, chaque modèle a été redimensionné pour s'inscrire dans une plage de valeurs allant de -1 à 1, tout en étant centré à l'origine. Nos choix de positionnement des modèles sont conformes aux conclusions énoncées dans [Blanz 99], qui indiquent que les vues préférées pour les objets familiers sont étroitement alignées avec le sens de gravité, offrant ainsi une perspective plus naturelle.

α_j	-$\pi/2$	-$\pi/4$	0	$\pi/4$	$\pi/2$
$label_\alpha(\alpha_j)$	*Dessous*	*Dessous*	$\emptyset$	*Dessus*	*Dessus*

β_i	$< \pi$			
$\beta_i[\pi]$	0	$\pi/4$	$\pi/2$	$3\pi/4$
$label_\beta(\beta_i)$	*Face*	*Face-Droit*	*Profil-Droit*	*Arrière-Droit*
β_i	$\geq \pi$			
$\beta_i[\pi]$	0	$\pi/4$	$\pi/2$	$3\pi/4$
$label_\beta(\beta_i)$	*Arriere*	*Arriere-Gauche*	*Profil-Gauche*	*Face-Gauche*

TABLE IV.4. – Labels des caméras en fonction de la valeur des angles (α_j, β_i).

Contrairement à certaines études utilisateurs et utilisatrices, comme celles réalisées dans [Secord 11] ou dans [Leifman 16], qui se basent sur des images, notre interface permet aux utilisateurs de manipuler des modèles 3D non texturés dans un environnement virtuel vide. Pour donner aux utilisateurs un large éventail de choix, nous leur offrons la possibilité de sélectionner leurs trois vues parmi 26 positions de caméra différentes, réparties sur une sphère centrée sur le modèle, comme illustré dans la Figure IV.18b. Cette stratégie de disposition minimise la redondance des points de vue et assure la diversité des angles de vue disponibles, permettant ainsi aux utilisateurs de sélectionner le point de vue le plus pertinent. Chaque caméra $C_{i,j}$ est définie par deux angles, notés α_j et β_i :

$\forall i \in [0,7],\ \forall j \in [0,4]$,

$$C_{i,j} = \begin{cases} x_{ij} = R\cos(\alpha_j)\cos(\beta_i) \\ y_{ij} = R\sin(\alpha_j) \\ z_{ij} = R\cos(\alpha_j)\sin(\beta_i) \end{cases} \tag{IV.6}$$

avec $R = 2.2$ le rayon de la sphère, $\beta_i = i.\pi/4$ et $\alpha_j = (2-j).\pi/4$.

Lorsque $|\alpha_j| = \pi/2 \Rightarrow \cos(\alpha_j) = 0$, nous obtenons huit caméras avec exactement la même position sur la sphère et donc offrant la même vue de l'objet sous huit rotations différentes. Dans le post-traitement, ces huit caméras seront considérées comme étant équivalentes puisqu'elles offrent exactement la même quantité d'information, à une rotation près.

Chacune des 26 caméras est associée à une étiquette spécifique qui dépend des valeurs de ses angles (α_j, β_i). Cette étiquette a pour but de faciliter l'identification et le traitement des points de vue en fournissant une indication visuelle immédiate de la position relative de la caméra par rapport à l'objet 3D. Par exemple, dans la Figure IV.18b, la caméra

rouge est étiquetée ***Dessus***, la caméra jaune est désignée comme ***Dessous***, tandis que la caméra bleu clair est identifiée comme ***Face***. Pour déterminer l'étiquette associée à une position de caméra particulière, il suffit de consulter la Table IV.4 que nous avons créée pour référencer ces associations et de respecter la condition suivante :
Soit une caméra $\mathbf{C}_{i,j}$, l'étiquette $e_{i,j}$ associée est définie par

$$e_{i,j} = \begin{cases} label_\alpha(\alpha_j) & \text{si } |\alpha_i| = \pi/2, \\ label_\alpha(\alpha_j) - label_\beta(\beta_i) & \text{sinon.} \end{cases} \tag{IV.7}$$

Notre répartition permet également la formation de groupes de caméras en fonction de leur position relative par rapport à l'objet 3D. Dans la Figure IV.18b, nous mettons en évidence ces sous-ensembles de caméras en utilisant une codification couleur distincte. Par exemple, le groupe de caméras de couleur violette est désigné sous l'appellation ***Milieu-Dessus*** et regroupe toutes les caméras positionnées de manière à offrir une perspective centrée et légèrement au-dessus de l'objet. Les caméras de couleur bleue, quant à elles, appartiennent au groupe ***Milieu*** et offrent des points de vue centrés à hauteur de l'objet. Enfin, les caméras vertes sont regroupées sous le nom ***Milieu-Dessous***, représentant les points de vue centrés et légèrement en dessous de l'objet.

8.5. Post-traitement des données récoltées

8.5.1. Réception des données

```
{'Tache N1': {
        'fichier_OBJ': 'horse_user_study.obj',
        'angle_initial_alpha': -0.785,
        'angle_initial_beta': 0.785,
        'choix_povs': [
                ['choix1', {'alpha': 0}, {'beta': 5.498}],
                ['choix2', {'alpha': 0}, {'beta': 4.712}],
                ['choix3', {'alpha': 0}, {'beta': 3.926}]]
}}
```

(a)

```
{'Analyse N1': {
        'nom_OBJ': 'horse',
        'mots_cles': [Recognizable, Pleasant, Side view]
}}
```

(b)

FIGURE IV.19. – Exemple des données enregistrées associées à la tâche n°1 dans (a) ou l'analyse n°1 (b), pour un utilisateur.

```
{'time': 1685949204700,
'type': 'start'},
...
{'type': 'fin inscription - choix tutorial'},
{'time': 1685949375231,
'type': 'debut tutorial'}
...
{'time': 1685949592393,
'type': 'debut choix vues'},
{'time': 1685949592393,
'type': 'debut tache N1'},
{'time': 1685949592441,
'type': 'T1 Choix_fait0 Affichage Mesh random : cap en alpha,
   beta : (-1.571, 4.712)'},
{'time': 1685949594178,
'type': 'T1 Choix_fait0 fleche droite (5, 4)'},
{'time': 1685949595208,
'type': 'T1 Choix_fait0 fleche haut (4, 3)'},
...
{'time': 1685949596608,
'type': 'T1 Choix_fait0 fleche gauche (5, 2)'},
...
{'time': 1685949599718,
'type': 'T1 Chox_fait1 choix N1 bouton pose :(7, 2)'},
...
{'time': 1685949613122,
'type': 'T1 Choix_fait3 bouton valider'},
{'time': 1685949613122,
'type': 'fin tache N1'},
{'time': 1685949613122,
'type': 'debut tache N2'},
...
```

FIGURE IV.20. – Extrait des interactions enregistrées dans un fichier JSON d'un utilisateur.

Dès qu'un utilisateur accède à l'étude en ligne, un fichier JSON est généré avec un identifiant unique, et il est progressivement complété au fur et à mesure que l'utilisateur progresse dans l'étude.

Dans un premier temps, nous enregistrons des données liées aux 10 modèles 3D rencontrés, notamment :

- le nom des modèles ;
- la position initiale des modèles définie par les deux angles (α_j, β_i) ;
- les coordonnées des trois choix effectués, dans l'ordre de sélection.

Ensuite, dans un deuxième temps, pour les 5 modèles parmi les 10 qui ont été analysés par les utilisateurs, nous enregistrons également les mots-clés choisis pour chaque modèle. Un exemple de données extraites d'un fichier JSON est illustré dans la Figure IV.19.

Chaque action effectuée par les utilisateurs, telle que les clics, est consignée dans leur fichier JSON respectif. Pour chaque interaction, nous enregistrons la date et attribuons une étiquette à l'action. Un extrait de ces interactions provenant d'un fichier JSON est présenté dans la Figure IV.20.

Enfin, des informations spécifiques aux utilisateurs, telles que leur pseudonyme, leur âge et leur nationalité, nous sont fournies par la plateforme Prolific. L'anonymat des utilisateurs est préservé, car chacun d'entre eux est associé à un identifiant unique par la plateforme.

8.5.2. Extraction d'informations statistiques

Au total, nous avons récolté 203 fichiers JSON. À partir de ces fichiers, nous pouvons extraire diverses statistiques, notamment la répartition des modèles 3D en fonction du nombre de fois qu'ils ont été traités ou analysés par les utilisateurs. Cette répartition est illustrée dans la Figure IV.21.

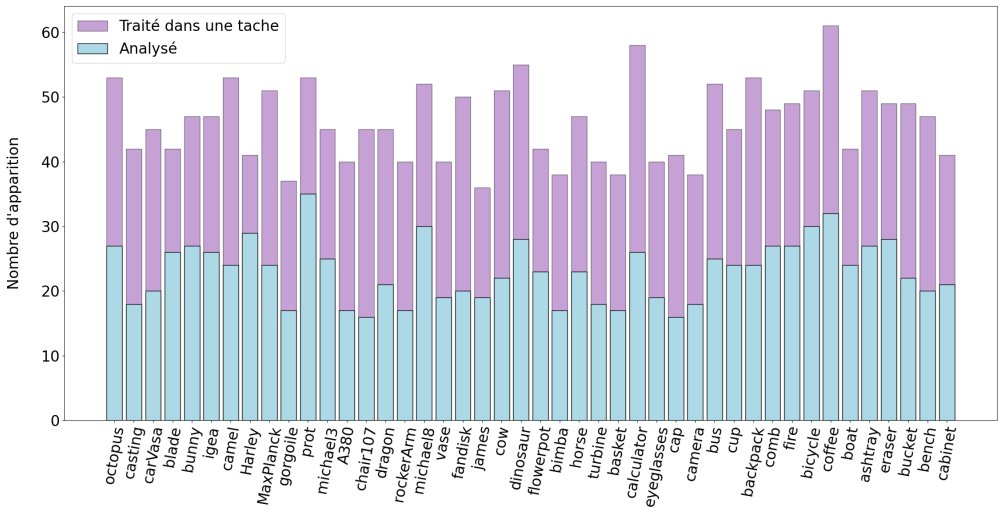


FIGURE IV.21. – Répartition des 44 modèles de l'étude qui ont été traités, puis, potentiellement, analysés.

Nous pouvons également extraire des données plus générales concernant les performances globales des utilisateurs. Ces données nous fournissent un aperçu complet de la manière dont les utilisateurs ont interagi avec notre interface et de leurs performances globales tout au long de l'étude. Voici un aperçu des données que nous avons extraites à partir des fichiers JSON relatifs aux utilisateurs :

- **Temps total de l'étude** : cette mesure représente la durée totale qu'un utilisateur a consacré à l'étude dans son ensemble, depuis le début jusqu'à la fin, cf. Figure IV.22b et Figure IV.22c.

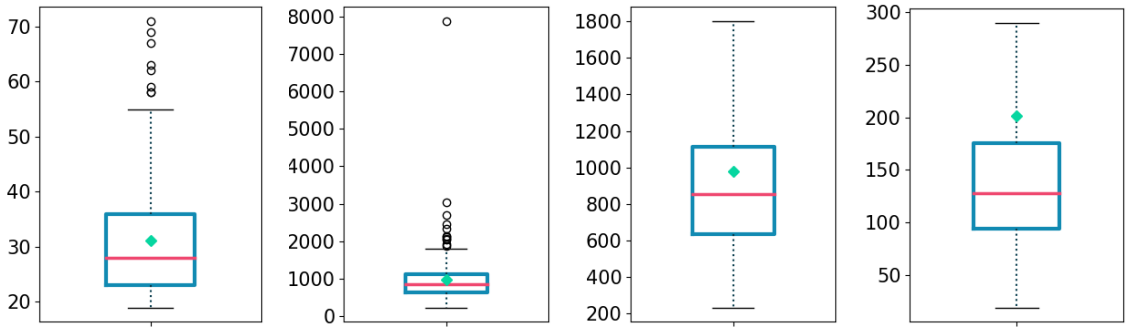
- **Temps passé sur chaque page** : nous avons enregistré la durée que les utilisateurs ont passé sur chaque page de l'étude, par exemple la page de description du contexte, la page d'inscription, etc. Cela nous permet de déterminer combien de temps, ils ont consacré à chaque aspect de l'étude. La répartition des temps passés sur la page du tutotiel sont disponibles dans la Figure IV.22d.
- **Temps moyen pour traiter chacun des modèles 3D** : nous avons sauvegardé le temps moyen que les utilisateurs ont consacré à évaluer les modèles 3D individuellement. Cette statistique nous donne un aperçu de la rapidité avec laquelle ils ont évalué chaque modèle, cf. Figure IV.22e.
- **Temps moyen pour analyser les modèles 3D** : cette mesure représente le temps moyen que les utilisateurs ont pris pour analyser et évaluer les cinq modèles 3D. Elle nous aide à comprendre le degré d'attention et d'analyse qu'ils ont consacré à chaque modèle, cf. Figure IV.22f.
- **Temps moyen avant de choisir le premier point de vue** : nous avons enregistré combien de temps les utilisateurs ont pris avant de sélectionner leur premier point de vue pour un modèle donné. Cela nous aide à évaluer leur processus de décision initial, cf. Figure IV.22g.
- **Pourcentage moyen de points de vue explorés** : nous avons calculé le pourcentage moyen des 26 points de vue disponibles que les utilisateurs ont explorés pour chaque modèle. Cette mesure nous indique comment ils ont examiné les diverses perspectives disponibles avant de faire leur choix, cf. Figure IV.22h.
- **Tutoriel complété** : nous avons utilisé une variable booléenne pour déterminer si un utilisateur a suivi l'intégralité du tutoriel proposé au début de l'étude. Cela nous permet de distinguer les utilisateurs qui ont eux une bonne compréhension de l'étude de ceux qui ont eu plus de problèmes.

8.5.3. Filtrage des participants

Suite aux observations réalisées dans [Nehmé 23], pour recruter des utilisateurs fiables, nous avons fait appel au service de *crowdsourcing* proposé par le site Prolific[5]. Grâce à cette plateforme, 203 utilisateurs ont complété notre étude : 121 hommes et 82 femmes. Ils proviennent de 23 pays différents et ont des âges compris entre 19 et 71 ans, cf. Figure IV.22a.

Afin de filtrer les utilisateurs avec les résultats les moins pertinents, nous avons utilisé trois paramètres différents, décrits précédemment, par utilisateur :

5. https://www.prolific.co/

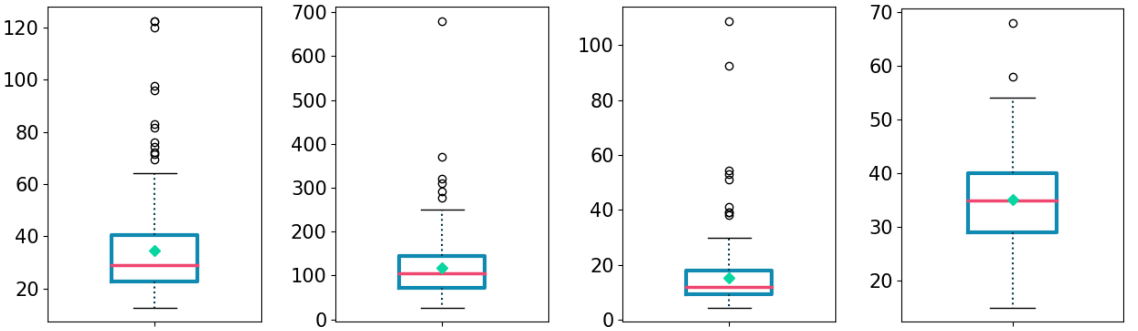


(a) Ages.

(b) Temps total (en sec) de l'étude.

(c) Temps total (en sec) de l'étude, sans valeurs abérrantes.

(d) Temps passé (en sec) sur la page du tutoriel.



(e) Temps moyen (en sec) pour traiter un objet 3D.

(f) Temps moyen (en sec) pour analyser les 5 modèles 3D.

(g) Temps moyen (en sec) avant de faire le premier choix.

(h) Pourcentage moyen des points de vue explorés.

FIGURE IV.22. – Données statistiques relatives aux 203 participants.

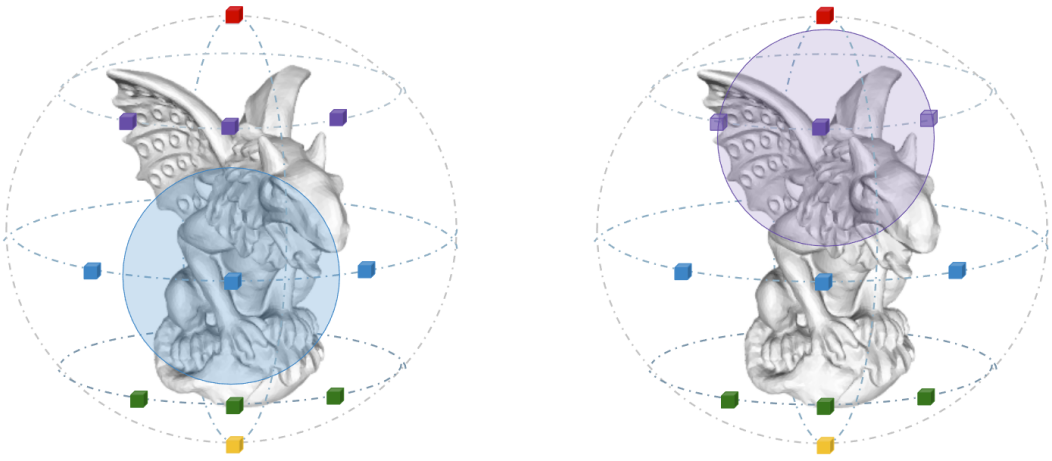
- Paramètre n°1 : le temps total pour compléter l'étude entière ;
- Paramètre n°2 : le temps total pour sélectionner les 3 points de vue pour l'ensemble des 10 modèles ;
- Paramètre n°3 : la navigation moyenne, autrement dit le pourcentage moyen de points de vue visités par modèles 3D.

Pour tous les participants, nous avons déterminé ces trois paramètres et avons calculé des données statistiques. En effet, pour chacun de ces trois paramètres p, nous avons calculé le premier quartile Q_1^p et l'interquartile IQR^p. Afin d'identifier les valeurs aberrantes, nous avons utilisé la méthode de l'écart interquartile pour établir une zone en dehors de Q_1^p. Si un utilisateur a l'un de ces trois paramètres plus petit que $Q_1^p - (1,5 * IQR^p)$, alors cet utilisateur est exclu. Selon ces critères, une seule utilisatrice a été retirée de l'étude.

8.5.4. Construction des histogrammes de popularité des points de vue

L'objectif est de déterminer quelles sont les vues les plus sélectionnées par les utilisateurs, autrement dit d'estimer la popularité de chaque vue. Pour ce faire, nous avons attribué un score à chacune des 26 caméras de l'étude pour chacun des 44 modèles étudiés, ce qui nous a permis de créer des histogrammes. Nous expliquons ci-dessous en détail comment nous avons procédé.

Comme mentionné précédemment, lors de l'étude utilisateur, il y a 26 caméras disponibles, notées POV_k, avec $k \in [1, 26]$. Dans la suite, nous considérons un modèle 3D M et un utilisateur u en particulier parmi les N utilisateurs qui ont traité ce modèle lors de leur expérience. Cet utilisateur doit choisir et classer trois points de vue parmi les 26 points de vue disponibles. Ces points de vue sélectionnés sont notés : $pov_{u,v} \in \{POV_k\}$, avec $v \in [1, 3]$. Chaque point de vue sélectionné $pov_{u,v}$ a un impact sur les points de vue de l'étude POV_k. En d'autres termes, chaque point de vue sélectionné $pov_{u,v}$ attribue un poids $p^k_{u,v}$ à chaque caméra POV_k. Ces poids sont calculés à l'aide d'une distribution normale 3D, inspirée de celle utilisée dans [Chave 15], centrée sur $pov_{u,v}$. Étant donné que la paramétrisation des positions des caméras n'est pas régulière dans l'étude, voir la Formule IV.6, un $pov_{u,v}$ peut avoir un impact sur plusieurs POV_k. En fonction de la position de la caméra sélectionnée, il existe deux configurations d'impact possibles, illustrées dans la Figure IV.23, En effet, nous avons estimé que les caméras étiquetées : ***Dessus***, ***Milieu*** et ***Dessous*** (voir la Figure IV.18b pour interpréter le code couleur) n'offrent pas des vues qui partagent suffisamment d'informations avec les vues des caméras voisines. De ce fait, ces caméras ne doivent pas avoir d'impact sur leurs caméras voisines, autrement dit, les valeurs des poids $p^k_{u,v}$ des caméras POV_k sont nulles. Cette configuration est illustrée dans la Figure IV.23a. Pour la deuxième situation, cela est différent.



(a) Aucun impact sur les caméras voisines (b) Impact sur les caméras voisines

FIGURE IV.23. – Deux configurations d'impact des caméras sélectionnées par les utilisateurs sur les caméras voisines en fonction de leurs positions sur le sphère.

En effet, les vues obtenues à partir des caméras étiquetées : ***Milieu-Dessus*** et ***Milieu-Dessous*** partagent des informations avec les vues offertes par les deux caméras voisines de même latitude. Un exemple de cette configuration est illustrée dans la Figure IV.23b. Les poids $p^k_{u,v}$ des deux caméras voisines ne doivent donc pas être nuls. Pour respecter ces conditions et avoir la quasi-totalité des valeurs qui se situent à moins de trois écarts-types de la moyenne, nous avons estimées, à l'aide des distances entre les caméras, la valeur de l'écart type à utiliser dans la distribution normale : $\sigma = 0.58$.

Chacun des choix effectués doit avoir la même importance les uns par rapport aux autres. Pour atteindre cet objectif, une normalisation est appliquée pour garantir que chaque choix $pov_{u,v}$ a la même importance. À cette fin, la somme des poids $p^k_{u,v}$ provenant de chaque $pov_{u,v}$ est normalisée à 1 :

$$\forall u \in [1, N], \forall v \in [1, 3] \sum_{k=1}^{26} p^k_{u,v} = 1$$

Enfin, pour le modèle M, nous souhaitons attribuer un score de popularité à chaque caméra POV_k, en prenant en compte l'ordre dans lequel les choix des points de vue ont été effectués. Pour calculer le score associé à la caméra POV_k, nous additionnons donc les poids $p^k_{u,v}$ associés à la caméra, provenant de tous les utilisateurs, en les pondérant par un facteur $(4 - v)$. Cette pondération signifie que les poids des points de vue choisis en premier $(v = 1)$ seront favorisés par rapport à ceux des points de vue choisis en dernier $(v = 3)$. La formule est la suivante :

$$\forall k \in [1, 26], \ score_k = \sum_{u=1}^{N} \sum_{v=1}^{3} (4 - v).p^k_{u,v} \tag{IV.8}$$

Après avoir traité les données extraites de l'étude utilisateur, nous avons interprété ces dernières pour saisir les tendances et les préférences émergentes des participants.

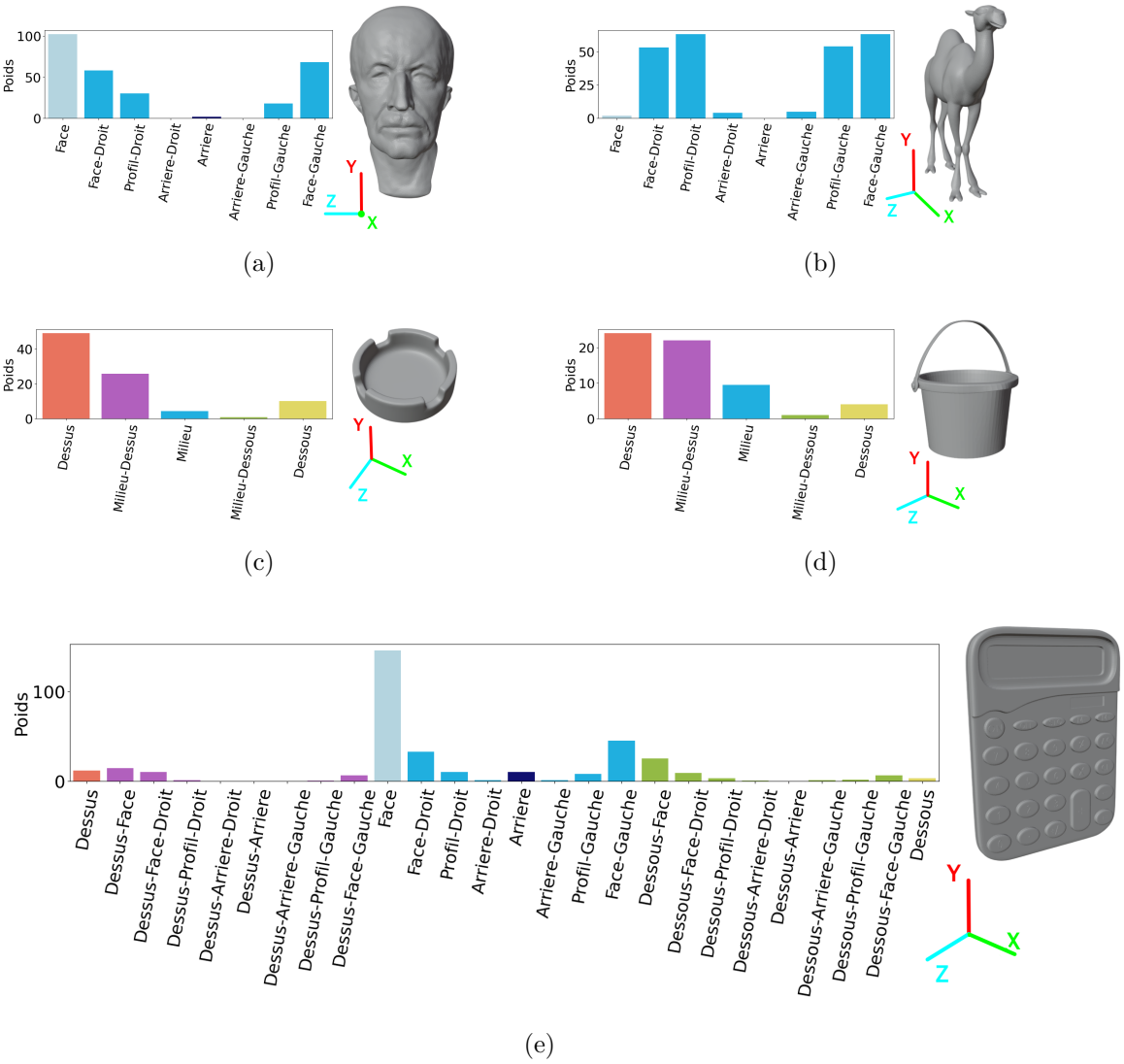
8.6. Analyses et interprétations des histogrammes

Une fois que nous avons récupéré les choix des utilisateurs et traité les données, nous sommes en mesure de générer des histogrammes qui illustrent la popularité de chacun des 26 points de vue parmi les utilisateurs. Ces histogrammes offrent un aperçu visuel de la distribution des préférences pour chaque point de vue, permettant ainsi une analyse approfondie de la sélection des utilisateurs.

8.6.1. Cas général

L'interprétation de ces histogrammes est un élément clé de notre étude, et elle peut être comparée aux conclusions émises dans [Blanz 99]. Contrairement à leur observation :

« *Most participants preferred off-axis views to straight front - or side-views* », nous constatons que cette préférence peut varier en fonction de la nature de l'objet examiné.



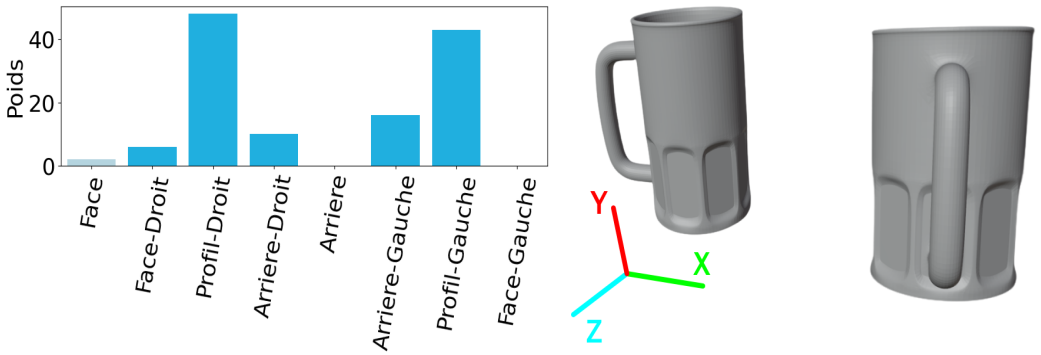
(a) (b) (c) (d) (e)

FIGURE IV.24. – Exemples de cinq histogrammes de popularité : dans (a) et (b), les points de vue représentés sont ceux étiquetés ***Milieu*** dans la Figure IV.18b, en (c) et (d) les points de vue sont regroupés en fonction de leur latitude, enfin en (e), tous les points de vue sont représentés.

Par exemple, pour les modèles représentant des êtres humains, les points de vue de ***Face*** sont largement privilégiés, comme illustré dans l'exemple de l'histogramme du buste de Max Planck dans la Figure IV.24a.
Cependant, pour les modèles d'animaux ou de créatures, nous observons que les vues obliques et latérales, regroupées sous les étiquettes {***Face-Droit***,***Profil-Droit***, ***Face-Gauche***, ***Profil-Gauche***}, sont fortement privilégiées, comme en témoigne l'histogramme du chameau présenté dans la Figure IV.24b.

En ce qui concerne les objets, la préférence des utilisateurs dépend largement de leur fonctionnalité, ce qui contredit les conclusions de [Blanz 99]. Par exemple, pour les objets qui servent de contenants, les points de vue ***Dessus*** et ***Milieu-Dessus*** sont massivement sélectionnés, comme l'illustrent les histogrammes des Figures IV.24c et IV.24d pour les modèles du saut et du cendrier. Enfin, pour des objets tels qu'une calculatrice, où les touches sont essentielles, la vue de ***Face*** est nettement prédominante, comme illustré dans la Figure IV.24e.

8.6.2. Vues accidentelles



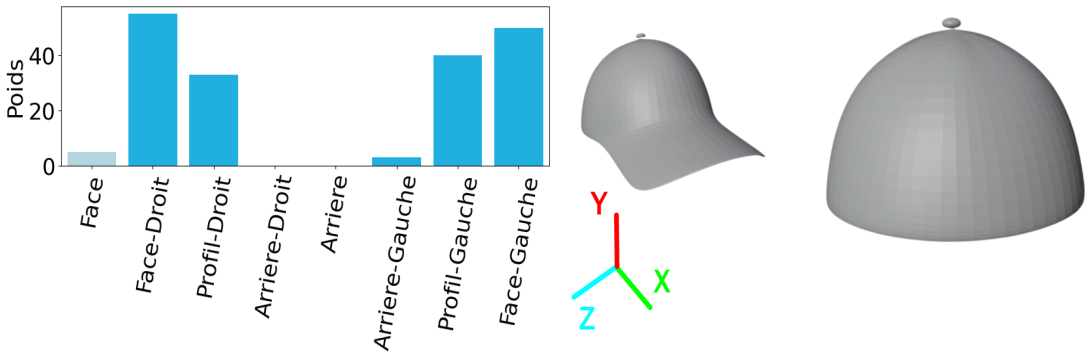
(a) Histogramme de popularité. (b) Vue ***Arrière***.

FIGURE IV.25. – L'histogramme de popularité proposé est celui associé au modèle 3D d'une tasse. Les points de vue représentés dans l'histogramme sont ceux étiquetés ***Milieu***, et la vue accidentelle de la tasse proposée en (b) est celle obtenue depuis la caméra ***Arrière***.

En accord avec les observations énoncées dans [Blanz 99], nos résultats montrent également une aversion pour les vues accidentelles parmi les utilisateurs. Les vues accidentelles désignent les points de vue qui posent des problèmes de perspective, autrement dit où une majeure partie des éléments de l'objet sont visibles mais avec une ambiguïté sur la compréhension. Par exemple, dans la Figure IV.25b, nous pouvons voir que la poignée de la tasse est alignée verticalement, créant une vue peu claire et potentiellement confuse. En conséquence, dans l'histogramme de la Figure IV.25a, la vue étiquetée ***Arrière***, qui présente cette ambiguïté, n'a jamais été choisie par les utilisateurs. Cette aversion pour les vues accidentelles confirme la préférence des utilisateurs pour des points de vue clairs et non ambigus, où la forme et la structure de l'objet sont facilement reconnaissables.

8.6.3. Vues occultées

Les points de vue où le modèle cache certaines de ses parties, comme dans le cas de la tasse qui cache sa poignée, sont également évités. Cette observation rejoint celle faite dans [Blanz 99]. Dans le cas de la tasse, le point de vue ***Face*** est très rarement choisi, comme l'illustre l'histogramme de la Figure IV.25a. Cela peut s'expliquer par le fait que, dans cette vue de face, la tasse ressemble davantage à un verre qu'à une tasse lorsque la anse n'est pas visible, créant ainsi une ambiguïté quant à la nature de l'objet. De manière similaire, pour la casquette, les utilisateurs n'ont jamais sélectionné le point de vue ***Arrière***, à partir duquel la visière n'est pas visible, comme le montre l'histogramme de la Figure IV.26a. Depuis ce point de vue, la casquette est difficilement identifiable, comme l'illustre la Figure IV.26b. Il est important de noter que les utilisateurs ont le temps, durant l'étude, de visionner les objets dans leur intégralité. Par conséquent, ils ont déjà identifié l'objet et en connaissent la nature avant de choisir leur point de vue.



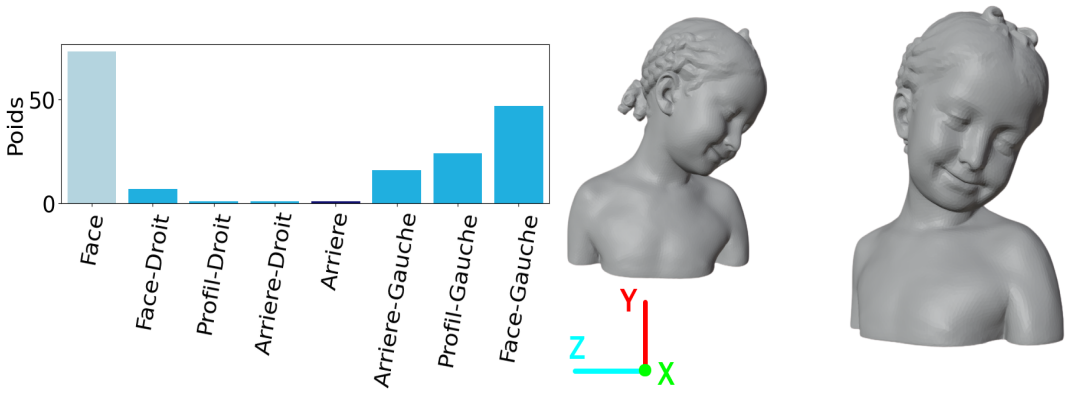
(a) Histogramme de popularité. (b) Vue ***Arrière***.

FIGURE IV.26. – L'histogramme de popularité proposé est celui associé au modèle 3D d'une casquette. Les points de vue représentés dans l'histogramme sont ceux étiquetés ***Milieu***, et la vue occultée de la casquette proposée en (b) est celle obtenue depuis la caméra ***Arrière***.

8.6.4. Objets avec des yeux

Selon les auteurs de [Secord 11] : « *When the object of interest is a creature with eyes or a face, we observe that people strongly prefer views where the eyes can be seen [Zusne 70]* ». Ainsi, dans sa formule, le poids associé à l'attribut qui quantifie la présence des yeux dans un point de vue est élevée. Cela signifie que les humains préfèrent voir des modèles dans les yeux, chaque fois que c'est possible. Cette hypothèse est confirmée par le fait que, comme mentionné précédemment, les utilisateurs choisissent des vues frontales lorsqu'ils observent un humain et des vues légèrement de profil pour les animaux et les créatures.

Prenons l'exemple du modèle de Bimba, sa particularité est qu'elle ne regarde pas droit devant, mais sur son côté inférieur gauche ; voir la Figure IV.27b. Selon l'histogramme, présenté dans la Figure IV.27a, le point de vue le plus sélectionné est ***Face***, mais les points de vue suivants les plus populaires sont ceux de son côté gauche.



(a) Histogramme de popularité. (b) Vue ***Face-Gauche***

FIGURE IV.27. – L'histogramme de popularité proposé est celui associé au modèle 3D de Bimba. Les points de vue représentés dans l'histogramme sont ceux étiquetés ***Milieu***, et la vue proposée en (b) est celle obtenue depuis la caméra ***Face-Gauche***.

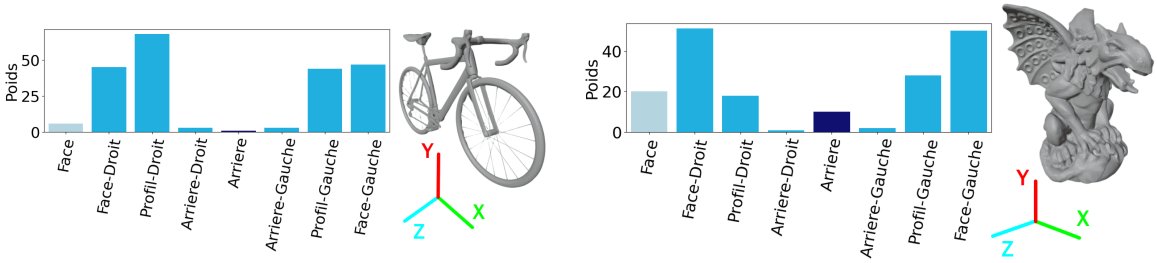
8.6.5. Objets symétriques

Pour les objets symétriques, les vues prises du côté gauche sont aussi souvent sélectionnées que la vue du côté droit, par rapport à l'objet. Cela peut être remarqué dans de nombreux histogrammes, tels que l'histogramme du vélo et l'histogramme de la gargouille, tous deux sur la Figure IV.28. Cette tendance est également visible avec des modèles 3D précédents, comme l'histogramme du chameau sur la Figure IV.24b, l'histogramme de la tasse affiché sur la Figure IV.25a et l'histogramme de la casquette illustré sur la Figure IV.26a

Cette indifférence entre les points de vue gauche ou droite peut indiquer que les informations sur l'objet sont équivalentes. Les utilisateurs choisissent probablement ces points de vue en fonction de leur propre préférence personnelle ou de leur habitude. Cela souligne l'importance de permettre aux utilisateurs de choisir parmi une variété de points de vue pour répondre à leurs préférences individuelles.

8.6.6. Objets non-familiers

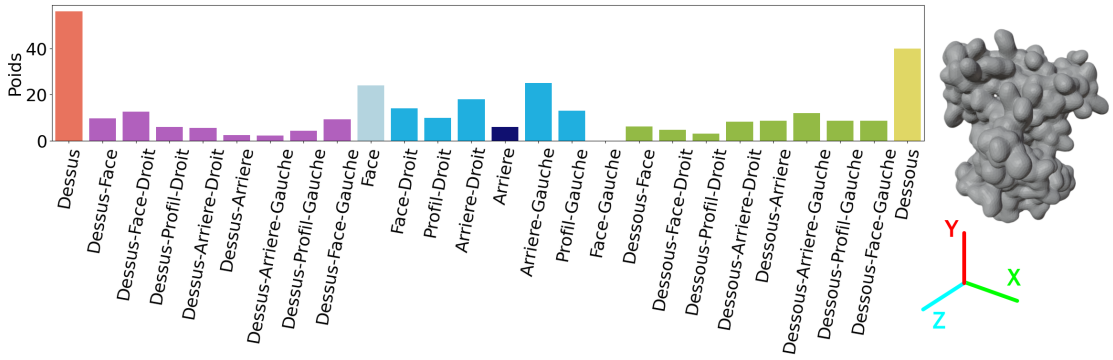
Nous avons choisi d'intégrer des modèles peu familiers dans notre étude : des pièces mécaniques et une molécule. Comme dans [Blanz 99], les nouveaux objets n'ont pas de



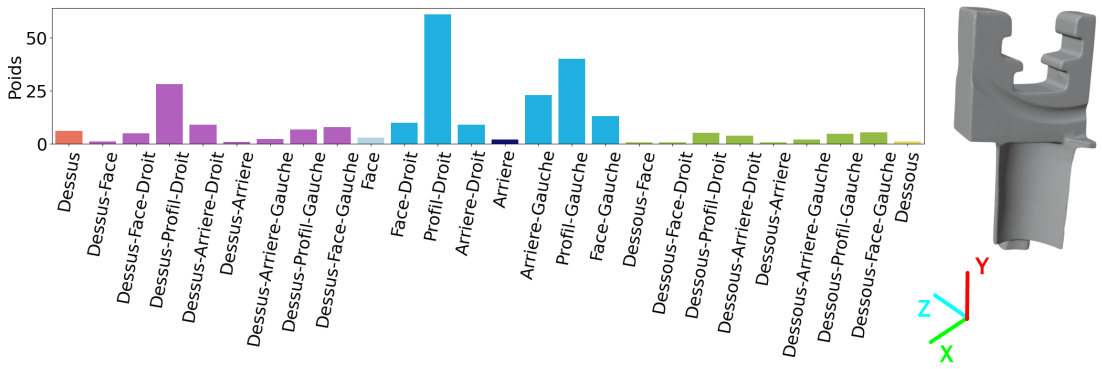
(a) Histogramme de popularité des vues d'un vélo.

(b) Histogramme de popularité des vues d'une gargouille.

FIGURE IV.28. – Les histogrammes de popularité proposés sont ceux associés aux modèles 3D d'objet symétrique : un vélo et une gargouille. Les points de vue représentés dans les histogrammes sont ceux étiquetés ***Milieu***.



(a) Histogramme de popularité des vues d'une molécule.



(b) Histogramme de popularité des vues d'une pièce mécanique.

FIGURE IV.29. – Les histogrammes de popularité proposés sont ceux associés aux modèles 3D d'objet non familiers : une molécule et une pièce mécanique. Les 26 points de vue étudiés sont tous représentés dans les histogrammes.

point de vue préféré : tous les points de vue ont été sélectionnés pour la molécule, comme le montre l'histogramme de la Figure IV.29a.

De plus, comme mentionné dans [Blanz 99], les points de vue accidentels sont toujours évités. Prenons l'exemple du modèle de la pièce mécanique, les points de vue qui donnent lieu à une ambiguïté, tels que ***Face*** ou ***Arrière*** ne sont pas beaucoup sélectionnés par les utilisateurs. Cette tendance suggère que, pour les objets moins familiers ou sans caractéristiques distinctes, les utilisateurs cherchent toujours des points de vue qui offrent une vue claire et non ambiguë, afin de mieux comprendre la nature et la forme de l'objet.

8.7. Comparaisons avec des images réelles

Nous avons cherché à observer les comportements des utilisateurs dans des conditions plus proches de la réalité. Contrairement à notre étude utilisateurs, où les objets étaient présentés sans textures dans un environnement neutre, nous avons envisagé l'impact d'agents extérieurs tels que la texture, l'éclairage ou la présence d'autres objets dans l'environnement sur le choix du meilleur point de vue. Nous avons donc voulu voir si ces agents avaient un réel impact sur ce choix. En dehors du cadre d'une étude où les utilisateurs sont contraints de choisir le point de vue le plus représentatif, la réalité diffère : les utilisateurs prennent des photos sans cette obligation. Dans ce contexte, il est probable que les choix de point de vue soient davantage influencés par l'esthétique que par la pertinence. Notre objectif est d'évaluer s'il existe une corrélation entre ces deux conditions de choix, mettant en lumière les éventuelles divergences entre le choix esthétique et le choix pertinent du point de vue.

Afin de découvrir s'il existe une corrélation entre ces deux conditions, nous avons utilisé la base de données *Objectnet3D* [Xiang 16] pour récupérer des images. Parmi nos 44 modèles 3D d'études, 22 d'entre eux appartiennent à une catégorie présente dans *Objectnet3D*. Cette base contient des milliers d'images pour une centaine de catégories, avec en moyenne une dizaine de modèles 3D par catégorie. Chaque image est annotée pour repositionner les différents objets dans des configurations similaires. Identifiant les catégories et modèles similaires aux nôtres, nous avons considéré chaque image comme représentant un point de vue choisi par un utilisateur. Contrairement à notre étude utilisateur où les participants devaient sélectionner trois points de vue, ici chaque image correspond à un unique point de vue.

Le traitement des données recueillies s'apparente à celui présenté dans la section 8.5. Une étape de mise en correspondance a été nécessaire pour comparer les points de vue issus de l'étude et ceux des images, ne correspondant pas toujours aux 26 points disponibles dans l'étude. Afin de ne pas perdre d'information, nous avons appliqué directement la distribution normale à partir de ces positions sur la sphère, calculant les poids d'impact sur les caméras voisines. Ensuite, comme dans le post-traitement de l'étude utilisateur,

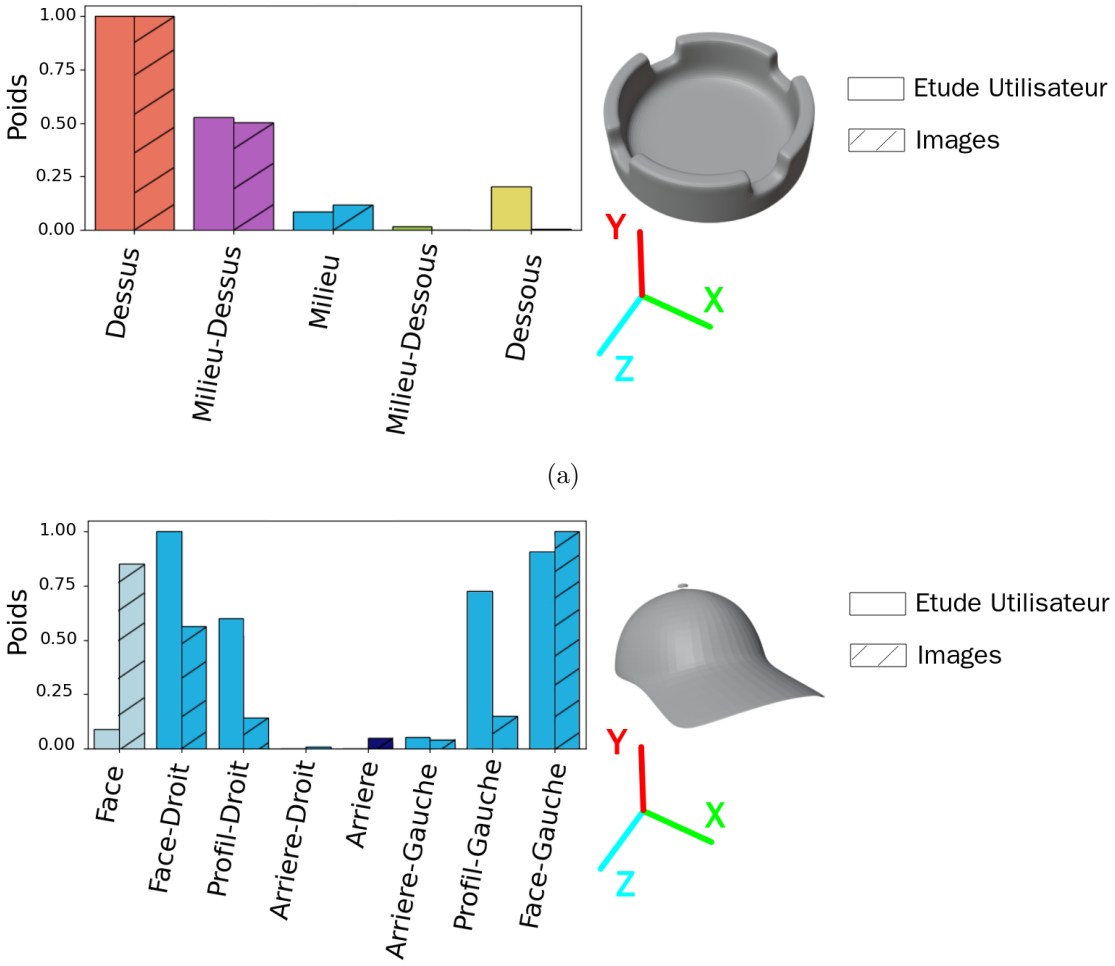
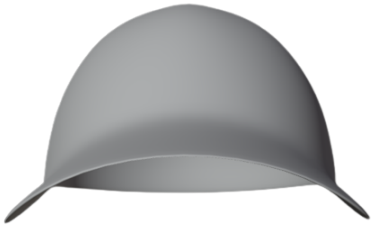


FIGURE IV.30. – Comparaisons des préférences humaines issues de l'étude utilisateur et extraites d'images réelles. En (a), les points de vue sont regroupés en fonction de leur latitude alors qu'en (b) les vues étiquetées ***Milieu***

nous avons calculé un histogramme par modèle 3D représentant la popularité des vues, issues des images, sur les 26 points de vue de l'étude. En procédant de la sorte, nous avons évalué quels étaient les points de vue les plus populaires lorsque les utilisateurs prenaient une photo dans des conditions réelles.

Nous avons analysé les meilleurs points de vue extraits de la collection d'images 2D et de l'étude utilisateur. Dans la plupart des cas, les points de vue les plus fréquemment utilisés pour prendre des photos sont similaires à ceux sélectionnés dans l'étude utilisateur. Par exemple, la Figure IV.30 montre la distribution des points de vue utilisés dans les deux situations : dans nos conditions expérimentales et dans des conditions réelles. Dans le cas du cendrier, les vues ***Dessus*** et ***Milieu-Dessus*** sont largement sélectionnées. Le principal facteur influençant le choix du point de vue est la fonctionnalité de l'objet. Cependant, le

cendrier est souvent un objet neutre en termes de texture.



(a) Vue de ***Face*** du modèle 3D.

(b) Image avec une vue de ***Face***

FIGURE IV.31. – Modèle 3D non texturé (a) et image (b) d'une casquette depuis le point de vue ***Face***.

Si nous prenons l'exemple de la casquette, nous constatons que les vues les plus populaires sont les vues ***Face-Droit*** et ***Face-Gauche***, voir la Figure IV.30b. La vue ***Arrière***, qui est une vue considérée comme occultée, est évitée dans les deux situations. Cependant, le point de vue ***Face*** est beaucoup plus choisi dans la vie réelle que dans l'étude. En effet, les logos sont souvent visibles sur le devant des casquettes, comme dans la Figure IV.31. Il y a une volonté de mettre en valeur le logo. Dans ce cas, il faut prendre en compte la texture et pas seulement la géométrie.

En résumé

Notre étude utilisateur se concentre sur la sélection de points de vue pour des modèles 3D, dans le but de comprendre les préférences des êtres humains.
L'interface que nous avons développée pour cette étude offre plusieurs fonctionnalités : elle permet aux personnes de manipuler des modèles 3D non texturés, et de suivre un tutoriel interactif pour comprendre l'utilisation de l'interface afin de sélectionner et de classer trois points de vue à partir de 26 options.
Nous avons recruté 203 personnes participantes, couvrant divers pays, avec une représentation femmes-hommes de 80%-20% environs et des âges allant de 19 à 71 ans. Pour garantir des données fiables, nous avons élaboré une étape de filtrage qui nous a permis de détecter et retirer une utilisatrice non convenable, d'après nos critères. Les données récoltées sont traitées pour créer des histogrammes, permettant d'évaluer la popularité des points de vue pour chacun des modèles 3D de notre base.
L'interprétation des résultats révèle que les préférences des personnes varient en fonction de la nature des modèles 3D. Nos analyses ont révélé une tendance à éviter les points de vue accidentels qui pourraient induire une ambiguïté dans l'identification des objets, une tendance qui se maintient même lors de l'étude d'objets moins familiers. De même, les vues contenant des éléments occultants, susceptibles de créer des confusions quant à la nature des objets, sont également évitées par les utilisatrices et utilisateurs. Nous avons constaté que les personnes préfèrent généralement les points de vue permettant de visualiser le visage, en particulier les yeux. De plus, pour la catégorie d'objets étudiée, nous avons observé une corrélation entre la fonctionnalité de l'objet observé et la popularité du point de vue. Ces observations restent valables même lors de l'analyse de photographies naturelles d'objets du quotidien. Nous avons observé que la fonctionnalité de l'objet continue d'influencer la décision concernant le point de vue, les utilisatrices et utilisateurs évitant toujours les vues accidentelles ou contenant des occultations. Toutefois, la texture de l'objet peut également exercer une influence sur les préférences des être humains.

Chapitre V

Conclusion générale

Contributions

La problématique de cette thèse correspond à la sélection automatique du point de vue 2D le plus pertinent pour un objet 3D donné, afin de garantir une identification sans ambiguïté de cet objet et une compréhension claire de ses caractéristiques essentielles et notamment de sa fonctionnalité. Pour résoudre cette problématique, deux axes de recherche ont été développés : un premier axe traitant le maillage de l'objet 3D avec un ensemble d'images de cet objet et un deuxième axe utilisant directement un maillage 3D de l'objet.

Dans le premier axe, étant donné un objet 3D, nous avons quantifié la pertinence des vues 2D en nous appuyant sur l'extraction de l'information essentielle à partir de ces deux modalités, en mettant particulièrement l'accent sur les caractéristiques géométriques de l'objet. Nous avons développé une méthode utilisant un détecteur multimodal de points saillants répétables utilisant la saillance curviligne multi-échelle, permettant d'ignorer les aspects esthétiques. Pour valider nos résultats, nous avons mis en place un protocole original impliquant la dégradation progressive d'images pour former notre vérité terrain.Une comparaison avec une méthode utilisant des scores de confiance issus des réseaux de neurones. Nous avons montré que l'approche déterministe, utilisant un score de pertinence, est nettement plus performante que celles s'appuyant sur le score de confiance de réseaux de neurones, et qu'elle est également la plus robuste. Le choix proposé de l'image la plus pertinente est donc automatique, et ne nécessite pas d'entraînement sur un grand ensemble de données. En utilisant cette quantification représentée par le score de pertinence, nous sommes en mesure de proposer des classements d'images et donc d'automatiser le choix de l'image d'un objet 3D la plus pertinente.

Dans le second axe, nous avons étudié la sélection du meilleur point de vue 2D d'un objet 3D directement à partir de son maillage 3D. Nous avons développé une méthode géométrique dont l'avantage réside dans la combinaison de la saillance intrinsèque de l'objet avec des attributs géométriques spécifiques à chaque point de vue, tels que la proportion de surface visible de l'objet ainsi que de ces yeux. Cette approche nous a permis de déterminer le point de vue le plus pertinent et représentatif d'un objet, en mettant en avant l'infor-

mation essentielle de celui-ci au travers de l'estimation de la saillance intrinsèque utilisant l'entropie des courbures locales. Pour valider nos résultats, nous avons réalisé une étude utilisateur et utilisatrice afin de comparer la méthode déterministe proposée avec d'autres approches de l'état de l'art. Les résultats ont confirmé la pertinence et l'efficacité de notre méthode pour sélectionner les meilleures vues 2D des objets 3D. En effet, comparée aux approches de référence, notre méthode sélectionne les points de vue les plus proches de la sélection des utilisateurs et utilisatrices. Notre analyse visuelle souligne également que lorsque notre approche diffère de l'étude, elle propose toujours un point de vue intéressant.

En conclusion, cette thèse a apporté des contributions à la résolution de la problématique de sélection des vues 2D pertinentes pour les objets 3D. Les deux axes de recherche développés ont permis de proposer des méthodes efficaces et validées à la fois de manière quantitative et qualitative.

Perspectives

Dans la continuité de nos travaux de thèse, plusieurs perspectives de recherche émergent pour approfondir et enrichir notre compréhension de l'estimation de la pertinence de vues 2D d'un objet 3D. Dans le cadre de notre premier axe d'étude portant sur la sélection d'images 2D représentant au mieux un objet 3D commun, une perspective serait d'impliquer des utilisateurs et utilisatrices dans ce processus de validation. Plutôt que d'utiliser uniquement les classements que nous élaborons, nous pourrions demander à des personnes de classer elles-mêmes les images en fonction de leur pertinence et ainsi obtenir des classements d'images plus précis et réalistes. Une autre piste intéressante serait d'explorer l'intégration de la sémantique dans le processus de sélection d'images. En effet, en plus des caractéristiques géométriques, l'inclusion d'informations sémantiques sur les objets pourrait améliorer la pertinence et la précision des images sélectionnées. Par exemple, pour départager des images de pertinence équivalente, une possibilité serait d'utiliser des techniques de reconnaissance d'objets dans les images 2D pour une sélection s'appuyant sur le contenu, tout en tenant compte de la signification et de la pertinence des objets eux-mêmes.

Pour notre deuxième axe d'étude, nous pourrions explorer l'impact de la texture et des couleurs des modèles 3D sur le processus de sélection de point de vue. Ces caractéristiques pourraient fournir des informations supplémentaires qui ne sont pas nécessairement visibles sur la structure des maillages 3D. En tenant compte de la texture et des couleurs, nous pourrions adapter notre approche de sélection de point de vue plus précise, cf. Figure V.1. De plus, pour optimiser l'efficacité de notre approche, nous pourrions envisager une présélection des vues à estimer en utilisant un module d'attention ou des cartes de fixations issues d'expériences de suivi du regard. De plus, en nous appuyant sur l'hypothèse que

FIGURE V.1. – Modèle 3D texturé. Il est pourvu de textures qui permettent de visualiser la position des yeux du canard, une caractéristique cruciale à considérer lors du choix du meilleur point de vue.

des objets de même classe ont souvent des zones d'attention similaires, nous pourrions compenser le manque de diversité de ces données en termes de modèles 3D utilisés.

Nous pourrions également explorer des méthodes utilisant des techniques d'apprentissage, y compris d'apprentissage profond, dans nos deux axes d'étude. Cependant, ces méthodes s'éloignent de notre objectif initial, qui est de développer des approches déterministes s'appuyant sur des informations caractéristiques géométriques et potentiellement sémantiques, extraites directement des objets 3D et/ou des images 2D. Notre priorité est de concevoir des méthodes interprétables facilement et robustes à toutes les données, ce qui peut être difficile à atteindre avec des méthodes d'apprentissage. En effet, bien que l'apprentissage profond puisse offrir des performances élevées dans certains cas, il peut également manquer de transparence.

Une autre perspective intéressante serait d'appliquer le concept d'information essentielle au domaine de la compression. Nous avons défini cette information comme l'ensemble compact des éléments caractéristiques et fondamentaux nécessaires à la définition et à la compréhension globale d'un objet 3D. Cette approche pourrait être pertinente en compression, où l'objectif est de déterminer les éléments indispensables à conserver et à transmettre d'un objet 3D. Ces éléments seraient ceux constituant l'information essentielle de l'objet 3D considéré. Cette sélection d'éléments caractéristique pourrait être similaire à celle proposée par [Judd 07], où les auteurs identifient, à partir d'une vue d'un modèle 3D, les caractéristiques à préserver sous forme de lignes pour créer une esquisse 2D du modèle. Appliquer le concept d'information essentielle à la compression permettrait d'optimiser le processus de transmission tout en préservant les caractéristiques essentielles de l'objet.

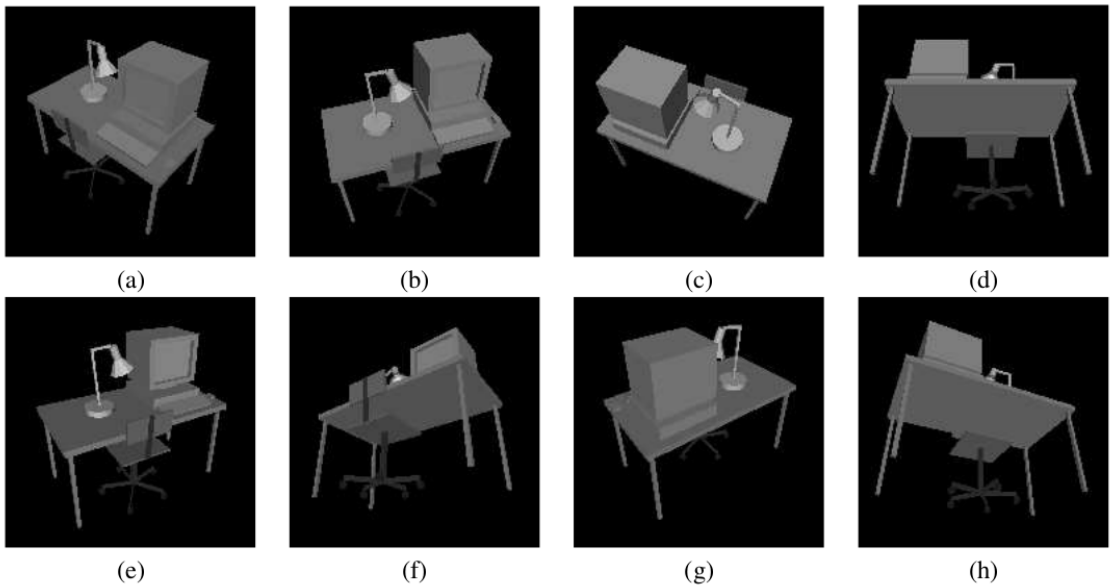


FIGURE V.2. – L'illustration extraite de [Vázquez 01] présente les huit meilleures vues pour une visualisation complète de la scène proposée. Le nombre de vues est un paramètre à spécifier en entrée de l'algorithme.

Enfin, une orientation vers des contextes dynamiques pourrait également être envisagée. Actuellement, nos axes d'études se concentrent sur la sélection des vues 2D les plus pertinentes pour un objet 3D statique. Toutefois, une extension dynamique consisterait à déterminer la vidéo la plus pertinente d'un objet 3D, constituée de l'ensemble des vues 2D les plus pertinentes de celui-ci. Cette approche dynamique nécessite de considérer plusieurs paramètres, tels que le nombre d'images dans la vidéo ou la durée totale. De plus, il est essentiel de garantir une certaine cohérence temporelle entre les transitions des points de vue, afin d'assurer la fluidité de la vidéo. Cette étude se rapprocherait plus du domaine de recherche de la compréhension de scène ou *Scene understanding*, cf. Figure V.2. Ce travail pourrait s'inspirer de la problématique définie dans [Zhao 14], où l'objectif est de définir la trajectoire d'une caméra autour d'un objet 3D pour visualiser l'ensemble de l'objet, tout en maximisant la redondance d'information entre deux points de vue consécutifs. Leur approche permet ainsi de minimiser le temps de téléchargement de l'information nécessaire pour compléter une nouvelle vue à partir de la précédente.

Dans cette perspective, l'objet 3D d'intérêt est statique tandis que la caméra est en mouvement ; une situation où la scène est aussi dynamique pourrait également être envisagée. Par exemple, nous pourrions avoir un point de vue fixe sur un objet évoluant dans le temps. L'idée serait alors de déterminer une séquence continue ou non des vues 2D les plus pertinentes pour visualiser l'évolution de l'objet. L'idée de sélectionner des vues essentielles sur l'échelle du temps trouve des parallèles avec le domaine de la bande dessinée, où quelques illustrations clés permettent aux dessinateurs de transmettre l'idée

(a) (b)

FIGURE V.3. – Illustrations de scènes dynamiques extraites de « l'art Invisible » [McCloud 93]. Les informations temporelles sont réduite à l'essentiel, ici avec seulement quelques vignettes et un texte court, deux actions complètes sont racontées.

du mouvement d'un personnage ou d'un objet dans une scène, comme le montre le Figure V.3 : on note la capacité, à la fois spatiale et temporelle, de la bande dessinée d'identifier une information essentielle à la compréhension de l'histoire.

Bibliographie

[Abdulwahab 19] S. Abdulwahab, H. A. Rashwan, J. Cristiano, S. Chambon et D. Puig. Effective 2D/3D Registration using Curvilinear Saliency Features and Multi-Class SVM. International Conference on Computer Vision Theory and Applications, 2019.

[Abid 20] M. Abid, M. Perreira Da Silva et P. Le Callet. Towards visual saliency computation on 3d graphical contents for interactive visualization. International Conference on Image Processing. IEEE, 2020.

[Achanta 12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua et S. Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. Transactions on pattern analysis and machine intelligence, 2012.

[Achlioptas 21] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny et L. J. Guibas. Artemis : Affective language for visual art. Conference on Computer Vision Pattern Recognition, 2021.

[Agarwal 11] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz et R. Szeliski. Building Rome in a day. Communications of the Association for Computing Machinery, ACM, 2011.

[Ahmed 18] S. M. Ahmed, Y. Z. Tan, C. M. Chew, A. A. Mamun et F. S. Wong. Edge and Corner Detection for Unorganized 3D Point Clouds with Application to Robotic Welding. International Conference on Intelligent Robots and Systems, 2018.

[Alcantarilla 12] P. F. Alcantarilla, A. Bartoli et A. Davison. KAZE features. European Conference on Computer Vision, ECCV. Springer, 2012.

[Alcantarilla 13] P. F. F Alcantarilla, J. Nuevo et A. Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. British Machine Vision Conference, BMVC, 2013.

[Aldana-Iuit 16] J. Aldana-Iuit, D. Mishkin, O. Chum et J. Matas. In the saddle : chasing fast and repeatable features. International Conference on Pattern Recognition, pages 675–680. IEEE, 2016.

[Altwaijry 16] H. Altwaijry, A. Veit et C. Belongie, S.and Tech. Learning to detect and match keypoints with deep architectures. British Machine Vision Conference, BMVC, 2016.

[Amirshahi 15] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler et C. Redies. Jenaesthetics subjective dataset : analyzing paintings by subjective scores. European Conference on Computer Vision, ECCV, 2015.

[Arvanitis 20] G. Arvanitis, A. S. Lalos et K. Moustakas. Robust and fast 3-D saliency mapping for industrial modeling applications. Transactions on Industrial Informatics, 2020.

[Back-face culling 24] Back-face culling. Back-face culling — Wikipedia, the free encyclopedia. 2024. [Online ; accessed 13-May-2024].

[Barroso-Laguna 22] A. Barroso-Laguna et K. Mikolajczyk. Key. net : Keypoint detection by handcrafted and learned cnn filters revisited. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[Barsalou 85] L. W. Barsalou. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of experimental psychology : learning, memory, and cognition, 11(4), 1985.

[Bay 08] H. Bay, A. Ess, T. Tuytelaars et L. Van Gool. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, CVIU, 2008.

[Beaudet 78] P. R. Beaudet. Rotationnally invariant image operators. International Conference on Pattern Recognition, 1978.

[Becker 07] M. W. Becker, H. Pashler et J. Lubin. Object-intrinsic oddities draw early saccades. Journal of Experimental Psychology : Human Perception and Performance, 2007.

[Ben Amor 10] M. Ben Amor, A. Samet, F. Kammoun et N. Masmoudi. Exploitation des caractéristiques du système visuel humain dans les métriques de qualité. Applications Médicales de l'Informatique : Nouvelles Approches, AMINA, 2010.

[Besl 92] P. J. Besl et N. D. McKay. A Method for Registration of 3-D Shapes. Transactions on pattern analysis and machine intelligence, 1992.

[Bhattacharya 10] S. Bhattacharya, R. Sukthankar et M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. ACM International Conference on Multimedia, ACMMM, 2010.

[Bhowmik 20] A. Bhowmik, S. Gumhold, C. Rother et E. Brachmann. Reinforced feature points : Optimizing feature detection and description for a high-level task. Conference on Computer Vision Pattern Recognition, 2020.

[Biglia 15] A. Biglia et P. Belleflamme. Analyse prospective sur l'implémentation de la voiture autonome : impact sur l'industrie automobile et le citoyen. Mémoire de master, Université Catholique de Louvain, Belgique, 2015.

[Biswas 23] S. Biswas, E. Kruijff et E. Veas. View recommendation for multi-camera demonstration-based training. Multimedia Tools and Applications, 2023.

[Blanz 99] V. Blanz et M. J Tarr. What object attributes determine canonical views ? Perception, 1999.

[Bo 11] L. Bo, X. Ren et D. Fox. Depth kernel descriptors for object recognition. International Conference on Intelligent Robots and Systems, 2011.

[Bonaventura 18] X. Bonaventura, M. Feixas, M. Sbert, L. Chuang et C. Wallraven. A survey of viewpoint selection methods for polygonal models. Entropy, 2018.

[Burbea 82] J. Burbea et C. Rao. On the convexity of some divergence measures based on entropy functions. Transactions on Information Theory, 1982.

[Bustos 09] B. Bustos et T. Schreck. Feature-Based 3D Object Retrieval. Springer, Boston, 2009.

[Bychkovsky 11] V. Bychkovsky, S. Paris, E. Chan et F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. Conference on Computer Vision Pattern Recognition. IEEE, 2011.

[Cai 04] D. Cai, X. He, Z. Li, W.-Y. Ma et J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. ACM International Conference on Multimedia, ACMMM, 2004.

[Camargo 09] J. E. Camargo et F. A. González. A multi-class kernel alignment method for image collection summarization. Iberoamerican Congress on Pattern Recognition, 2009.

[Campbell 01] R. J. Campbell et P. J. Flynn Flynn. A Survey Of Free-Form Object Representation and Recognition Techniques. Computer Vision and Image Understanding, CVIU, 2001.

[Castelein 23] W. Castelein, Z. Tian, T. Mchedlidze et A. Telea. Based Quality for Analyzing and Exploring 3D Multidimensional Projections. International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2023.

[Cetinic 19] E. Cetinic, T. Lipic et S. Grgic. A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access, 2019.

[Chambah 03] M. Chambah, A. Rizzi, C. Gatta, B. Besserer et D. Marini. Perceptual approach for unsupervised digital color restoration of cinematographic archives. Color Imaging VIII : Processing, Hardcopy, and Applications, volume 5008. International Society for Optics and Photonics, 2003.

[Chambah 07] M. Chambah, A. Rizzi et C. Saint Jean. Image quality and automatic color equalization. Image Quality and System Performance, IQSP. International Society for Optics and Photonics, 2007.

[Chambon 11] S. Chambon et A. Crouzil. Occlusions handling in dense stereo matching. Pattern Recognition Letters, PR, 44(9) :2063–2075, 2011.

[Chambon 20] S. Chambon. Analyse de contenus visuels complémentaires statiques et dynamiques : Détection de saillance et de similarités en 2D et en 3D pour reconnaitre des objets en identifiant forme et apparence. Habilitation à diriger des recherches, Université de Toulouse, Institut National Polytechnique de Toulouse, INPT, 2020.

[Chang 16] H. Chang, F. Yu, J. Wang, D. Ashley et A. Finkelstein. Automatic triage for a photo series. ACM transactions on graphics, TOG, 35, 2016.

[Chave 15] A. D Chave. A note about Gaussian statistics on a sphere. Geophysical Journal International, 2015.

[Chen 00] J.-H. Chen, C. A Bouman et J.-C. Dalton. Hierarchical browsing and search of large image databases. Transactions on Image Processing, TIP, 2000.

[Chen 12] X. Chen, A. Saparov, B. Pang et T. Funkhouser. Schelling points on 3D surface meshes. ACM transactions on graphics, TOG, 2012.

[Choy 15] C. B. Choy, M. Stark, S. Corbett-Davies et S. Savarese. Enriching object detection with 2D-3D registration and continuous viewpoint estimation. Conference on Computer Vision Pattern Recognition, 2015.

[Collins 20] T. Collins, D. Pizarro, S. Gasparini, N. Bourdel, P. Chauvet, M. Canis, L. Calvet et A. Bartoli. Augmented Reality Guided Laparoscopic Surgery of the Uterus. Transactions on Medical Imaging, TMI, 40(1) :371–380, 2020.

[Cornia 20] M. Cornia, M. Stefanini, L. Baraldi et R. Cucchiara. Meshed-memory transformer for image captioning. Conference on Computer Vision Pattern Recognition, 2020.

[Crombez 18] N. Crombez, R. Seulin, O. Morel, D. Fofi et C. Demonceaux. Multimodal 2D image to 3D model registration via a mutual alignment of sparse and dense visual features. IEEE international conference on Robotics and Automation, ICRA, 2018.

[Cyr 00] C. M. Cyr, A. F. Kamal, T. B. Sebastian et B. B. Kimia. 2D-3D Registration Based on Shape Matching. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA, pages 198–203, 2000.

[Dalal 05] N. Dalal et B. Triggs. Histograms of oriented gradients for human detection. Conference on Computer Vision Pattern Recognition, 2005.

[Darom 12] T. Darom et Y. Keller. Scale-Invariant Features for 3-D Mesh Models. Transactions on Image Processing, TIP, 2012.

[Datta 06] R. Datta, D. Joshi, J. Li et J. Z. Wang. Studying aesthetics in photographic images using a computational approach. European Conference on Computer Vision, ECCV. Springer, 2006.

[Daubechies 92] I. Daubechies. Ten lectures on wavelets. SIAM, 1992.

[Dellepiane 07] M. Dellepiane, M. Callieri, F. Ponchio et R. Scopigno. Mapping highly detailed color information on extremely dense 3D models : the case of David's restoration. Eurographic, 2007.

[Deng 07a] D. Deng. Content-based image collection summarization and comparison using self-organizing maps. Pattern recognition, 2007.

[Deng 07b] H. Deng, W. Zhang, E. Mortensen, T. Dietterich et L. Shapiro. Principal Curvature-Based Region Detector for Object Recognition. Conference on Computer Vision Pattern Recognition, 2007.

[Deng 09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei. Imagenet : A large-scale hierarchical image database. Conference on Computer Vision Pattern Recognition, 2009.

[Deng 17] Y. Deng, C. C. Loy et X. Tang. Image aesthetic assessment : An experimental survey. Signal Processing Magazine, 2017.

[DeTone 18] D. DeTone, T. Malisiewicz et A. Rabinovich. Superpoint : Self-supervised interest point detection and description. Conference on Computer Vision Pattern Recognition, 2018.

[Dhar 11] S. Dhar, V. Ordonez et T Berg. High level describable attributes for predicting aesthetics and interestingness. Conference on Computer Vision Pattern Recognition, 2011.

[Ding 19] X. Ding, W. Lin, Z. Chen et X. Zhang. Point cloud saliency detection by local and global feature fusion. Transactions on Image Processing, TIP, 2019.

[Ding 23] X. Ding, Z. Chen, W. Lin et Z. Chen. Towards 3D Colored Mesh Saliency : Database and Benchmarks. Transactions on Multimedia, 2023.

[dos Anjos 23] R. K. dos Anjos, R. A. Roberts, B. Allen, J. Jorge et K. Anjyo. Saliency detection for large-scale mesh decimation. Computers & Graphics, 2023.

[Dosovitskiy 20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et others. An image is worth 16x16 words : Transformers for image recognition at scale. Conference on Learning Representations, 2020.

[Drucker 03] S. Drucker, C. Wong, A. Roseway, S. Glenner et S. De Mar. Photo-triage : Rapidly annotating your digital photographs. Rapport Technique, Microsoft, 2003.

[Dubey 15] R. Dubey, J. Peterson, A. Khosla, M. H. Yang et B. Ghanem. What makes an object memorable ? Conference on Computer Vision Pattern Recognition, 2015.

[Dufek 21] J. Dufek, X. Xiao et R. R. Murphy. Best viewpoints for external robots or sensors assisting other robots. Transactions on Human-Machine Systems, 2021.

[Dutagaci 10] H. Dutagaci, C. P. Cheung et A. Godil. A benchmark for best view selection of 3D objects. ACM workshop on 3D object retrieval, 2010.

[Fabricatore 02] C. Fabricatore, M. Nussbaum et R. Rosas. Playability in action videogames : A qualitative design model. Human-Computer Interaction, 2002.

[Fang 17] Y. Fang, J. Yan, L. Li, J. Wu et W. Lin. No reference quality assessment for screen content images with both local and global feature representation. Transactions on Image Processing, TIP, 2017.

[Feixas 09] M. Feixas, M. Sbert et F. González. A unified information-theoretic framework for viewpoint selection and mesh saliency. ACM Transactions on Applied Perception, 2009.

[Fischer 14] P. Fischer et T. Brox. Image Descriptors Based on Curvature Histograms. German Conference on Pattern Recognition, GCPR, 2014.

[Flitton 10] G. Flitton, T. Breckon et N. Megherbi Bouallagu. Object Recognition using 3D SIFT in Complex CT volumes. British Machine Vision Conference, BMVC, 2010.

[Foley 94] J Foley, S Feiner et J Hughes. Computer Graphics : Principle and Practice. Addions Wesley, 1994.

[Förstner 87] W. Förstner et E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. ISPRS Intercommission conference on fast processing of photogrammetric data, 1987.

[Gal 06] R. Gal et D. Cohen-Or. Salient geometric features for partial shape matching and similarity. ACM transactions on graphics, TOG, 2006.

[Gales 11] G. Gales. Mise en correspondance de pixels pour la stéréovision binoculaire par propagation d'appariements de points d'intérêt et sondage de régions. PhD thesis, Université de Toulouse, 2011.

[Gao 05] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng et W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. ACM International Conference on Multimedia, ACMMM, 2005.

[Geirhos 17] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge et F. A. Wichmann. Comparing deep neural networks against humans : object recognition when the signal gets weaker. CoRR (arXiv), 2017.

[Geirhos 18] R. Geirhos, C. Temme, J. Rauber, H. Schütt, M. Bethge et F. Wichmann. Generalisation in humans and deep neural networks. Advances in neural information processing systems, NIPS, 2018.

[Geirhos 19] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann et W. Brendel. ImageNet-trained CNNs are biased towards texture ; increasing shape bias improves accuracy and robustness. CoRR (arXiv), 2019.

[Geirhos 21] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. Wichmann et W. Brendel. Partial success in closing the gap between human and machine vision. Advances in neural information processing systems, NIPS, 2021.

[Girshick 14] R. Girshick, J. Donahue, T. Darrell et J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. Conference on Computer Vision Pattern Recognition, 2014.

[Girshick 15] R. Girshick. Fast r-cnn. Conference on Computer Vision Pattern Recognition, 2015.

[Gu 14] M. Gu, S. Hu, X. Wang, X. Liang, X. Shen et A. Qin. Saliency-driven Depth Compression for 3D Image Warping. PG (Short Papers), 2014.

[Guo 12] Y. Guo, M. Liu, T. Gu et W. Wang. Improving photo composition elegantly : Considering image similarity during composition optimization. Computer Graphics Forum, 2012.

[Guo 20] C. Guo, C. Li, J. Guo, C. Loy, J. Hou, S. Kwong et R. Cong. Zero-reference deep curve estimation for low-light image enhancement. Conference on Computer Vision Pattern Recognition, 2020.

[Habibi 15] Z. Habibi, E. M. Mouaddib et G. Caron. Good feature for framing : Saliency-based Gaussian mixture. International Conference on Intelligent Robots and Systems, pages 3682–3687. IEEE, 2015.

[Hafiz 15] D. A. Hafiz, B. A. B. Youssef, W. M. Sheta et H. A. Hassan. Interest Point Detection in 3D Point Cloud Data Using 3D Sobel-Harris Operator. International Journal of Pattern Recognition and Artificial Intelligence, PRAI, 2015.

[Hall 05] P. Hall et M. Owen. Simple Canonical Views. British Machine Vision Conference, BMVC, 2005.

[Harris 88] C. Harris et M. Stephens. A combined corner and edge detector. Alvey Vision Conference, AVC, 1988.

[Hartwig 22] S. Hartwig, M. Schelling, C. Onzenoodt, P.-P. Vázquez, P. Hermosilla et T. Ropinski. Learning human viewpoint preferences from sparsely annotated models. Computer Graphics Forum, 2022.

[He 16] K. He, X. Zhang, S. Ren et J. Sun. Deep residual learning for image recognition. Conference on Computer Vision Pattern Recognition, 2016.

[He 17] K. He, G. Gkioxari, P. Dollár et R. Girshick. Mask R-CNN. Conference on Computer Vision Pattern Recognition, 2017.

[He 22] S. He, Y. Zhang, R. Xie, D. Jiang et A. Ming. Rethinking image aesthetics assessment : Models, datasets and benchmarks. International Joint Conference on Artificial Intelligence, 2022.

[Hosu 20] V. Hosu, H. Lin, T. Sziranyi et D. Saupe. KonIQ-10k : An ecologically valid database for deep learning of blind image quality assessment. Transactions on Image Processing, TIP, 2020.

[Hu 20] S. Hu, X. Liang, H. Shum, F. Li et N. Aslam. Sparse metric-based mesh saliency. Neurocomputing, 2020.

[Huang 17] G. Huang, Z. Liu, L. Van Der Maaten et K. Weinberger. Densely connected convolutional networks. Conference on Computer Vision Pattern Recognition, 2017.

[Huang 20] J. Huang, C. Cui, C. Zhang, Z. Shen, J. Yu et Y. Yin. Learning multi-scale attentive features for series photo selection. International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020.

[Inoue 18] N. Inoue, R. Furuta, T. Yamasaki et K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. Conference on Computer Vision Pattern Recognition, 2018.

[Irschara 09] A. Irschara, C. Zach, J. M. Frahm et H. Bischof. From structure-from-motion point clouds to fast location recognition. Conference on Computer Vision Pattern Recognition, 2009.

[Isola 11] P. Isola, D. Parikh, A. Torralba et A. Oliva. Understanding the intrinsic memorability of images. Advances in neural information processing systems, NIPS, 2011.

[Jacobs 10] D. E Jacobs, D. B. Goldman et E. Shechtman. Cosaliency : Where people look when comparing images. ACM symposium on User interface software and technology, UIST, pages 219–228, 2010.

[Jaffe 06] A. Jaffe, M. Naaman, T. Tassa et M. Davis. Generating summaries for large collections of geo-referenced photographs. Proceedings of the 15th international conference on World Wide Web, WWW, pages 853–854, 2006.

[Jang 21] H. Jang et J.-S. Lee. Analysis of deep features for image aesthetic assessment. IEEE Access, 2021.

[Janoch 13] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko et T. Darrell. A category-level 3d object dataset : Putting the kinect to work. Consumer depth cameras for computer vision. Springer, 2013.

[Jeong 17] S.-W. Jeong et J.-Y. Sim. Saliency detection for 3D surface geometry using semi-regular meshes. Transactions on Multimedia, 2017.

[Jiang 15] M. Jiang, S. Huang, J. Duan et Q. Zhao. Salicon : Saliency in context. Conference on Computer Vision Pattern Recognition, CVPR, 2015.

[Jiang 18] M. Jiang, Y. Wu, T. Zhao, Z. Zhao et C. Lu. Pointsift : A sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv :1807.00652, 2018.

[Jin 19] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou et X. Zhou. Aesthetic attributes assessment of images. ACM International Conference on Multimedia, ACMMM, 2019.

[Judd 07] T. Judd, F. Durand et E. H. Adelson. Apparent ridges for line drawing. ACM transactions on graphics, TOG, 2007.

[Judd 09] T. Judd, K. Ehinger, F. Durand et A. Torralba. Learning to predict where humans look. International Conference on Computer Vision, ICCV, 2009.

[Kamgar-Parsi 11] B. Kamgar-Parsi et B. Kamgar-Parsi. Matching 2D image lines to 3D models : Two improvements and a new algorithm. Conference on Computer Vision Pattern Recognition, 2011.

[Kao 17] Y. Kao, R. He et K. Huang. Deep aesthetic quality assessment with semantic information. Transactions on Image Processing, TIP, 2017.

[Kaufman 12] L. Kaufman, D. Lischinski et M. Werman. Content-Aware Automatic Photo Enhancement. Computer Graphics Forum, 2012.

[Kennedy 08] L. S Kennedy et M. Naaman. Generating diverse and representative image search results for landmarks. Proceedings of the 15th international conference on World Wide Web, WWW, 2008.

[Khosla 14] A. Khosla, A. Das Sarma et R. Hamid. What makes an image popular ? Proceedings of the 15th international conference on World Wide Web, WWW, 2014.

[Kim 17] S.-H. Kim, Y.-W. Tai, J.-Y. Lee, J. Park et I. S. Kweon. Category-Specific Salient View Selection via Deep Convolutional Neural Networks. Computer Graphics Forum, 2017.

[Kim 19] J. H. Kim et Y. Kim. Instagram user characteristics and the color of their photos : Colorfulness, color diversity, and color harmony. Information Processing & Management, 2019.

[Kim 21] S. Kim, M. Jeong et B. Ko. Self-supervised keypoint detection based on multi-layer random forest regressor. Transactions on pattern analysis and machine intelligence, 2021.

[Kitchen 82] L. Kitchen et A. Rosenfeld. Gray level corner detection. Pattern Recognition Letters, PRL, 1982.

[Knoop 09] S. Knoop, S. Vacek et R. Dillmann. Fusion of 2D and 3D sensor data for articulated body tracking. Robotics and Autonomous Systems, 2009.

[Knopp 10] J. Knopp, M. Prasad, G. Willems, R. Timofte et L. Van Gool. Hough transform and 3D SURF for robust three dimensional classification. European Conference on Computer Vision, ECCV, 2010.

[Komorowski 18] J. Komorowski, K. Czarnota, T. Trzcinski, L. Dabala et S. Lynen. Interest point detectors stability evaluation on ApolloScape dataset. European Conference on Computer Vision, ECCV, 2018.

[Kong 16] S. Kong, X. Shen, Z. Lin, R. Mech et C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. European Conference on Computer Vision, ECCV. Springer, 2016.

[Kozegar 16] E. Kozegar. Rule Of Photography In Image Saliency Detection. Conference on Knowledge-Based Engineering and Innovation, KBEI, 2016.

[Krizhevsky 12] A. Krizhevsky, I. Sutskever et G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, NIPS, 25, 2012.

[Kubilius 16] J. Kubilius, S. Bracci et H. P. Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. PLoS computational biology, 12(4), 2016.

[Kuzovkin 17] D. Kuzovkin, T. Pouli, R. Cozot, O. Le Meur, J. Kervec et K. Bouatouch. Context-aware clustering and assessment of photo collections. Symposium on Computational Aesthetics, 2017.

[Kuzovkin 18] D. Kuzovkin, T. Pouli, R. Cozot, O. Le Meur, J. Kervec et K. Bouatouch. Image selection in photo albums. ACM International Conference on Multimedia Retrieval, ICMR, 2018.

[Kwon 20] B. Kwon, J. Huh, K. Lee et S. Lee. Optimal camera point selection toward the most preferable view of 3-d human pose. IEEE Transactions on Systems, Man, and Cybernetics : Systems, 2020.

[Laga 10] H. Laga. Semantics-driven approach for automatic selection of best views of 3D shapes. Eurographics, 2010.

[Lai 11] K. Lai, L. Bo, X. Ren et D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. International conference on robotics and automation, 2011.

[Lake 15] B. M. Lake, W. Zaremba, R. Fergus et T. M. Gureckis. Deep Neural Networks Predict Category Typicality Ratings for Images. Cognitive Science, CogSci, 2015.

[Lau 16] M. Lau, K. Dev, W. Shi, J. Dorsey et H. Rushmeier. Tactile mesh saliency. ACM transactions on graphics, TOG, 2016.

[Lavoué 18] G. Lavoué, F. Cordier, H. Seo et M.-C. Larabi. Visual attention for rendered 3D shapes. Computer Graphics Forum, 2018.

[LeCun 98] Y. LeCun, L. Bottou, Y. Bengio et P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.

[Leder 04] H. Leder, B. Belke, A. Oeberst et D. Augustin. A model of aesthetic appreciation and aesthetic judgments. British journal of psychology, 2004.

[Leder 22] H. Leder, J. Hakala, V.-T. Peltoketo, C. Valuch et M. Pelowski. Swipes and saves : A taxonomy of factors influencing aesthetic assessments and perceived beauty of mobile phone photographs. Frontiers in Psychology, 2022.

[Lee 05] C. H. Lee, A. Varshney et D. W Jacobs. Mesh saliency. ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH, 2005.

[Lee 07] Y. Lee, L. Markosian, S. Lee et J. F. Hughes. Line Drawings via Abstracted Shading. ACM transactions on graphics, TOG, 2007.

[Lee 11] T.-Y. Lee, O. Mishchenko, H.-W. Shen et R. Crawfis. View point evaluation and streamline filtering for flow visualization. IEEE Pacific Visualization Symposium, 2011.

[Lee 13] Y. Y. Lee, M. K. Park, J. D. Yoo et K. H. Lee. Multi-Scale Feature Matching between 2D Image and 3D Model. ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH, 2013.

[Lei 17] H. Lei, G. Jiang et L. Quan. Fast Descriptors and Correspondence Propagation for Robust Global Point Cloud Registration. Transactions on Image Processing, TIP, 2017.

[Leifman 16] G. Leifman, E. Shtrom et A. Tal. Surface regions of interest for viewpoint selection. Transactions on pattern analysis and machine intelligence, 2016.

[Li 09] C. Li et T. Chen. Aesthetic visual quality assessment of paintings. IEEE Journal of selected topics in Signal Processing, 2009.

[Li 19] J. Li et G. H. Lee. Usip : Unsupervised stable interest point detection from 3d point clouds. International Conference on Computer Vision, ICCV, 2019.

[Li 23] B. Li, Y. Cheng et W. Li. Keypoint Detection Based on Curvature Grouping and Adaptive Sampling. Available at SSRN 4608637, 2023.

[Limper 16] M. Limper, A. Kuijper et D. W Fellner. Mesh Saliency Analysis via Local Curvature Entropy. Eurographics, 2016.

[Lin 14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick et P. Dollár. Microsoft COCO : Common Objects in Context. CoRR (arXiv), abs/1405.0312, 2014.

[Liu 05] L. Liu et I. Stamos. Automatic 3D to 2D registration for the photorealistic rendering of urban scenes. Conference on Computer Vision Pattern Recognition, 2005.

[Liu 10] L. Liu, R. Chen, L. Wolf et D. Cohen-Or. Optimizing photo composition. Computer Graphics Forum, 2010.

[Liu 12] H. Liu, L. Zhang et H. Huang. Web-image driven best views of 3D shapes. The Visual Computer, 2012.

[Liu 16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu et A. C. Berg. SSD : Single Shot MultiBox Detector. Lecture Notes in Computer Science, LNCS, 2016.

[Liu 21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin et B. Guo. Swin transformer : Hierarchical vision transformer using shifted windows. CoRR (arXiv), 2021.

[Lo 12] L.-Y. Lo et J.-C. Chen. A statistic approach for photo quality assessment. International Conference on Information Security and Intelligent Control. IEEE, 2012.

[Lowe 04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal on Computer Vision, IJCV, 2004.

[Lu 14] X. Lu, Z. Lin, H. Jin, J. Yang et J. Z. Wang. Rapid : Rating pictorial aesthetics using deep learning. ACM International Conference on Multimedia, ACMMM, 2014.

[Lu 15] X. Lu, Z. Lin, X. Shen, R. Mech et J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. International Conference on Computer Vision, ICCV, pages 990–998, 2015.

[Lu 20] Y. Lu, C. Guo, Y. Lin, F. Zhuo et F.Y. Wang. Computational aesthetics of fine art paintings : The state of the art and outlook. Acta Automatica Sinica, 2020.

[Luo 00] M. R. Luo, G. Cui et B. Rigg. Derivation of a rotation fuction for the new cie colour difference formula. colour and visual scales, 2000.

[Luo 08] Y. Luo et X. Tang. Photo and video quality evaluation : Focusing on the subject. European Conference on Computer Vision, ECCV, 2008.

[Luo 20] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang et L. Quan. Aslfeat : Learning local features of accurate shape and localization. Conference on Computer Vision Pattern Recognition, 2020.

[Ma 19] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang et S. Wang. Image and video compression with neural networks : A review. IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[Mainali 13] P. Mainali, G. Lafruit, Q. Yang, B. Geelen, L. V. Gool et R. Lauwereins. SIFER : scale-invariant feature detector with error resilience. International journal of computer vision, 2013.

[Mair 10] E. Mair, G. Hager, D. Burschka, M. Suppa et G. Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. European Conference on Computer Vision, ECCV. Springer, 2010.

[Malon 18] T. Malon, P. Guyot, G. Roman-Jimenez, S. Chambon, V. Charvillat, A. Crouzil, A. Péninou, J. Pinquier, F. Sèdes et C. Sénac. Toulouse campus surveillance dataset : scenarios, soundtracks, synchronized videos with overlapping and disjoint views. ACM Multimedia Systems Conference, MMSys, 2018.

[Marsaglia 21] N. Marsaglia, Y. Kawakami, S. D. Schwartz, S. Fields et H. Childs. An entropy-based approach for identifying user-preferred camera positions. Symposium on Large Data Analysis and Visualization, 2021.

[Matas 02] J. Matas, O. Chum, U. Martin et T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. British Machine Vision Conference, BMVC, 2002.

[McCloud 93] S. McCloud. Understanding Comics: The invisible Art. Harper Collins, 1993.

[McDonnell 09] R. McDonnell, M. Larkin, B. Hernández, I. Rudomin et C. O'Sullivan. Eye-catching crowds : saliency based selective variation. ACM transactions on graphics, TOG, 2009.

[Meierhold 10] N. Meierhold, M. Spehr, A. Schilling, S. Gumhold et H. G. Maas. Automatic feature matching between digital images and 2D representations of a 3D laser scanner point cloud. International archives of the photogrammetry, remote sensing and spatial information sciences, ISPRS, 2010.

[Mezuman 12] E. Mezuman et Y. Weiss. Learning about canonical views from internet image collections. Advances in neural information processing systems, NIPS, 2012.

[Miao 10] Y Miao et J. Feng. Perceptual-saliency extremum lines for 3D shape illustration. The Visual Computer, 2010.

[Mikolajczyk 04] K. Mikolajczyk et C. Schmid. Scale & affine invariant interest point detectors. International Journal on Computer Vision, IJCV, 2004.

[Mohammadi 23] S. Mohammadi et J. Ascenso. Predictive Sampling for Efficient Pairwise Subjective Image Quality Assessment. ACM International Conference on Multimedia, ACMMM, 2023.

[Mokhtarian 03] F. Mokhtarian, M. Bober, F. Mokhtarian et M. Bober. Robust image corner detection through curvature scale space. Transactions on pattern analysis and machine intelligence, 2003.

[Moravec 80] H. Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. Stanford University, 1980.

[More 21] J. More, D. Sutar, R. Sequeira et V. Chavan. Eye Detection using Haar Cascade Classifier. International Research Journal of Engineering and Technology, 2021.

[Morel 09a] J.-M. Morel et G. Yu. ASIFT : A new framework for fully affine invariant image comparison. SIAM journal on imaging sciences, 2009.

[Morel 09b] J.-M. Morel et G. Yu. ASIFT : A new framework for fully affine invariant image comparison. SIAM journal on imaging sciences, 2009.

[Mortara 09] M. Mortara et M. Spagnuolo. Semantics-driven best view of 3D shapes. Computers & Graphics, 2009.

[Murray 12] N. Murray, L. Marchesotti et F. Perronnin. AVA : A large-scale database for aesthetic visual analysis. Conference on Computer Vision Pattern Recognition, 2012.

[Muttenthaler 22] L. Muttenthaler, J. Dippel, L. Linhardt, R. A Vandermeulen et S. Kornblith. Human alignment of neural network representations. arXiv preprint arXiv :2211.01201, 2022.

[Nader 14] G. Nader, G. Guennebaud et N. Mellado. Adaptive multi-scale analysis for point-based surface editing. Computer Graphics Forum, 2014.

[Nam 11] W. H. Nam, D.-G. Kang, D. Lee, J. Y. Lee et J. B. Ra. Automatic registration between 3D intra-operative ultrasound and pre-operative CT images of the liver based on robust edge matching. The international journal of biomedical physics and engineering, Physics in Medicine & Biology, PMB, 2011.

[Nanda 21] A. Nanda, B. B. Mohapatra, APK Mahapatra, APK Abiresh Prasad Kumar Mahapatra et APK Mahapatra. Multiple comparison test by Tukey's honestly significant difference (HSD) : Do the confident level control type I error. IJAMS, pages 59–65, 2021.

[Naraoka 07] R. Naraoka. Locating an optimal light source for volume rendering. IIEEJ Image Electronics and Visual Computing Workshop 2007 (DVD), 2007.

[Nehmé 23] Y. Nehmé, J. Delanoy, F. Dupont, J.-P. Farrugia, P. Le Callet et G. Lavoué. Textured mesh quality assessment : Large-scale dataset and deep learning-based quality metric. ACM transactions on graphics, TOG, 2023.

[Nishiyama 11] M. Nishiyama, T. Okabe, I. Sato et Y. Sato. Aesthetic quality classification of photographs based on color harmony. Conference on Computer Vision Pattern Recognition. IEEE, 2011.

[Nouri 15] A. Nouri, C. Charrier et O. Lézoray. Multi-scale mesh saliency with local adaptive patches for viewpoint selection. Signal Processing : Image Communication, 2015.

[Ohtake 04] Y. Ohtake, A. Belyaev et H.-P. Seidel. Ridge-valley Lines on Meshes via Implicit Surface Fitting. ACM transactions on graphics, TOG, 2004.

[Ono 18] Y. Ono, E. Trulls, P. Fua et K. Yi. LF-Net : Learning local features from images. Advances in neural information processing systems, NIPS, 2018.

[Ouni 09] S. Ouni, E. Zagrouba, M. Chambah et M. Herbin. Vers une métrique de description objective d'une sensation subjective. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 2009.

[Page 03] D. Page, A. Koschan, S. Sukumar, B. Roui-Abidi et M. Abidi. Shape analysis algorithm based on information theory. International Conference on Image Processing. IEEE, 2003.

[Pan 19] B. Pan, S. Wang et QiQ.sheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. Conference on Artificial Intelligence, 2019.

[Park 12] J. Park, J-K. Lee, Y.-W. Tai et I. Kweon. Modeling photo composition and its application to photo re-arrangement. International Conference on Image Processing, 2012.

[Pasini 22] A. Pasini, F. Giobergia, E. Pastor et E. Baralis. Semantic Image Collection Summarization With Frequent Subgraph Mining. IEEE Access, 2022.

[Paudel 14] D. P. Paudel, C. Demonceaux, A. Habed et P. Vasseur. Localization of 2D Cameras in a Known Environment Using Direct 2D-3D Registration. International Conference on Pattern Recognition, 2014.

[Pelissier-Combescure 21] M. Pelissier-Combescure, G. Morin et S. Chambon. Extraction et comparaison d'information saillante : Pose favorable et image 2D révélatrice d'un objet 3D. ORASIS, 2021. in french.

[Pelissier-Combescure 22] M. Pelissier-Combescure, G. Morin et S. Chambon. Quelle image met le mieux en valeur un modèle 3D ? Congrès Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP 2022). AFRIF (Association Française pour la Reconnaissance et l'Interprétation des Formes), 2022.

[Pelissier-Combescure 23] M. Pelissier-Combescure, G. Morin et S. Chambon. To Quantify an Image Relevance Relative to a Target 3D Object. Scandinavian Conference on Image Analysis. Springer, 2023.

[Pelissier-Combescure 24] Marie Pelissier-Combescure, Sylvie Chambon et Géraldine Morin. Most Relevant Viewpoint of an Object : A View-Dependent 3D Saliency Approach. International Conference on Computer Vision Theory and Applications, 2024.

[Peng 16] K.-C. Peng et T. Chen. Toward correlating and solving abstract tasks using convolutional neural networks. Conference on Applications of Computer Vision, WACV, 2016.

[Plemenos 96] D. Plemenos et M. Benayada. Intelligent display in scene modelling. New techniques to automatically compute good views. International Conference GraphiCon, 1996.

[Plemenos 04] D. Plemenos, M. Sbert et M. Feixas. On viewpoint complexity of 3D scenes. International Conference GraphiCon, 2004.

[Plemenos 06] D. Plemenos et D. Sokolov. Intelligent scene display and exploration. International Conference GraphiCon, 2006.

[Plötz 17] T. Plötz et S. Roth. Automatic Registration of Images to Untextured Geometry Using Average Shading Gradients. International Journal on Computer Vision, IJCV, 2017.

[Podolak 06] J. Podolak, P. Shilane, A. Golovinskiy, S. Rusinkiewicz et T. Funkhouser. A planar-reflective symmetry transform for 3D shapes. ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH, 2006.

[Polonsky 05] O. Polonsky, G. Patané, S. Biasotti, C. Gotsman et M. Spagnuolo. What's in an image ? Towards the computation of the "best" view of an object. The Visual Computer, 2005.

[Pomerleau 15] F. Pomerleau, F. Colas et R. Siegwart. A Review of Point Cloud Registration Algorithms for Mobile Robotics. Foundations and Trends in Robotics, 2015.

[Prashnani 18] E. Prashnani, H. Cai, Y. Mostofi et P. Sen. Pieapp : Perceptual image-error assessment through pairwise preference. Conference on Computer Vision Pattern Recognition, 2018.

[Rashwan 19] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin et V. Charvillat. Using Curvilinear Features in Focus for Registering a Single Image to a 3D Object. Transactions on Image Processing, TIP, 2019.

[Redmon 16] J. Redmon, S. Divvala, R. Girshick et A. Farhadi. You only look once : Unified, real-time object detection. Conference on Computer Vision Pattern Recognition, 2016. The last version of Yolo is available on this link.

[Ren 15] S. Ren, K. He, R. Girshick et J. Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. Advances in neural information processing systems, NIPS, 28, 2015.

[Ren 17] J. Ren, X. Shen, Z. Lin, R. Mech et D. Foran. Personalized image aesthetics. International Conference on Computer Vision, ICCV, 2017.

[Ren 20] J. Ren, X. Shen, Z. Lin et R. Mech. Best frame selection in a short video. Conference on Applications of Computer Vision, WACV, 2020.

[Riahi Samani 20] Z. Riahi Samani et M. Ebrahimi Moghaddam. Image Collection Summarization Method Based on Semantic Hierarchies. Conference on Artificial Intelligence, 2020.

[Ricard 05] J. Ricard. Indexation et recherche d'objets 3D à partir de requêtes 2D et 3D. Thèse de doctorat, Université Claude Bernard Lyon 1, 2005.

[Rosenhahn 04] B. Rosenhahn, H. Ho et R. Klette. Block matching based 2D-3D pose estimation. Image and vision computing, IVC, 2004.

[Rosten 06] E. Rosten et T. Drummond. Machine Learning for High-Speed Corner Detection. European Conference on Computer Vision, ECCV, 2006.

[Rosten 08] E. Rosten, R. Porter et T. Drummond. Faster and better : A machine learning approach to corner detection. Transactions on pattern analysis and machine intelligence, 2008.

[Rudoy 12] D. Rudoy et L. Zelnik-Manor. Viewpoint selection for human actions. International Journal on Computer Vision, IJCV, 97(3) :243–254, 2012.

[Saha 15] A. Saha et Q. M. J. Wu. Utilizing image scales towards totally training free blind image quality assessment. Transactions on Image Processing, TIP, 2015.

[Sartori 15] A. Sartori, V. Yanulevskaya, A. A. Salah, J. Uijlings, E. Bruni et N. Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. ACM Transactions on Interactive Intelligent Systems, 2015.

[Sattler 11] T. Sattler, B. Leibe et L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. International Conference on Computer Vision, ICCV, 2011.

[Savinov 17] N. Savinov, A. Seki, L. Ladicky, T. Sattler et M. Pollefeys. Quad-networks : unsupervised learning to rank for interest point detection. Conference on Computer Vision Pattern Recognition, 2017.

[Sbert 05] M. Sbert, D. Plemenos, , M. Feixas et F. Gonzalez. Viewpoint Quality : Measures and Applications. Computational Aesthetics in Graphics, Visualization and Imaging, 2005.

[Scharstein 02] D. Scharstein et R. Szeliski. A Taxomomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal on Computer Vision, IJCV, 2002.

[Schelling 21] M. Schelling, P. Hermosilla, P.-P. Vázquez et T. Ropinski. Enabling viewpoint learning through dynamic label generation. Computer Graphics Forum. Wiley Online Library, 2021.

[Schwarz 18] K. Schwarz, P. Wieschollek et H. Lensch. Will people like your image ? learning the aesthetic space. Conference on Applications of Computer Vision, WACV, 2018.

[Scovanner 07] P. Scovanner, S. Ali et M. Shah. A 3-dimensional sift descriptor and its application to action recognition. ACM International Conference on Multimedia, ACMMM, 2007.

[Secord 11] A. Secord, J. Lu, A. Finkelstein, M. Singh et A. Nealen. Perceptual models of viewpoint preference. ACM transactions on graphics, TOG, 2011.

[Serre 19] T. Serre. Deep learning : the good, the bad, and the ugly. Annual review of vision science, 2019.

[Shaiek 12] A. Shaiek et F. Moutarde. 3D keypoints detection for objects recognition. International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV, 2012.

[Shannon 48] C. E. Shannon. A mathematical theory of communication. The Bell system technical journal, 1948.

[Sharma 15] V. Sharma, N. Kumar, Aand Agrawal, P. Singh et R. Kulshreshtha. Image summarization using topic modelling. International Conference on Signal and Image Processing Applications (ICSIPA), 2015.

[Sharma 22] D. Kumar Sharma, A. Singh, S. K. Sharma, G. Srivastava et J. Lin. Task-specific image summaries using semantic information and self-supervision. Soft Computing, 2022.

[Shen 19] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng et Z. He. RF-Net : An end-to-end image matching network based on receptive field. Conference on Computer Vision Pattern Recognition, 2019.

[Sheng 18] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang et B.-G. Hu. Attention-based multi-patch aggregation for image aesthetic assessment. ACM International Conference on Multimedia, ACMMM, 2018.

[Sheng 20] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, B.-G. Hu, R. Ji et C. Ma. Revisiting image aesthetic assessment via self-supervised feature learning. Conference on Artificial Intelligence, 2020.

[Simon 07] I. Simon, N. Snavely et S. M. Seitz. Scene summarization for online image collections. International Conference on Computer Vision, ICCV, 2007.

[Simonyan 14] K. Simonyan et A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR (arXiv), 2014.

[Singh 19] A. Singh, L. Virmani et A. Subramanyam. Image corpus representative summarization. International Conference on Multimedia Big Data, 2019.

[Singh 20] A. Singh et D. K. Sharma. Image collection summarization : Past, present and future. Data Visualization and Knowledge Engineering : Spotting Data Points with Artificial Intelligence, 2020.

[Sinha 11] P. Sinha, S. Mehrotra et R. Jain. Summarization of personal photologs using multidimensional content and context. ACM International Conference on Multimedia Retrieval, ICMR, 2011.

[Sipiran 11] I. Sipiran et B. Bustos. Harris 3D : a robust extension of the Harris operator for interest point detection on 3D meshes. The Visual Computer, 2011.

[Sipiran 13] I. Sipiran et B. Bustos. Key-components : detection of salient regions on 3D meshes. The Visual Computer, 2013.

[Smith 97] S. Smith et J. Brady. SUSAN – A New Approach to Low Level Image Processing. International Journal on Computer Vision, IJCV, 23(1) :45–78, 1997.

[Sokolov 05] D. Sokolov et D. Plemenos. Viewpoint quality and scene understanding. Eurographics Symposium on Virtual Reality, 2005.

[Song 12a] R. Song, Y. Liu, R. R. Martin et P. L. Rosin. Saliency-guided integration of multiple scans. Conference on Computer Vision Pattern Recognition, CVPR, 2012.

[Song 12b] R. Song, Y. Liu, Y. Zhao, R. R Martin et P. L Rosin. Conditional random field-based mesh saliency. International Conference on Image Processing. IEEE, 2012.

[Song 14] R. Song, Y. Liu, R. R Martin et P. L Rosin. Mesh saliency via spectral processing. ACM transactions on graphics, TOG, 2014.

[Song 19] R. Song, Y. Liu et P. L. Rosin. Mesh saliency via weakly supervised classification-for-saliency CNN. Transactions on visualization and computer graphics, 2019.

[Song 21] R. Song, W. Zhang, Y. Zhao, Y. Liu et P. L Rosin. Mesh saliency : An independent perceptual measure or a derivative of image saliency ? Conference on Computer Vision Pattern Recognition, CVPR, 2021.

[Song 23] R. Song, W. Zhang, Y. Zhao, Y. Liu et P. L. Rosin. 3D Visual saliency : an independent perceptual measure or a derivative of 2d image saliency ? Transactions on pattern analysis and machine intelligence, 2023.

[Spicer 19] J. Spicer et A. N. Sanborn. What does the mind learn ? A comparison of human and machine learning representations. Current opinion in neurobiology, 55, 2019.

[Sreelakshmi 21] PR. Sreelakshmi et S. Manmadhan. Image summarization using unsupervised learning. International conference on advanced computing and communication systems, 2021.

[Stan 03] D. Stan et I. K. Sethi. eID : A system for exploration of image databases. Information processing & management, 2003.

[Stoev 02] S. L Stoev et W. Strasser. A case study on automatic camera placement and motion for visualizing historical data. Visualization, 2002.

[Strecha 09] C. Strecha, A. Lindner, K. Ali et P Fua. Training for task specific keypoint detection. German Conference on Pattern Recognition, GCPR. Springer, 2009.

[Streiff 21] D. Streiff, L. Bernreiter, F. Tschopp, Marius Fehr et R. Siegwart. 3D3L : Deep learned 3D keypoint detection and description for lidars. IEEE international conference on Robotics and Automation, ICRA, 2021.

[Su 21] Y.-H. Su, K. Huang et B. Hannaford. Multicamera 3d viewpoint adjustment for robotic surgery via deep reinforcement learning. Journal of Medical Robotics Research, 2021.

[Sun 18] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B Tenenbaum et W. T. Freeman. Pix3D : Dataset and methods for single-image 3D shape modeling. Conference on Computer Vision Pattern Recognition, 2018.

[Szegedy 15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke et A. Rabinovich. Going deeper with convolutions. Conference on Computer Vision Pattern Recognition, 2015.

[Tamaazousti 11] M. Tamaazousti, V. Gay-Bellile, S. N. Collette, S. Bourgeois et M. Dhome. NonLinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. Conference on Computer Vision Pattern Recognition, 2011.

[Tan 19] M. Tan et Q. Le. Efficientnet : Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, ICML, 2019.

[Tan 21] M. Tan et Q. V. Le. Efficientnetv2 : Smaller models and faster training. International Conference on Machine Learning, ICML, 2021.

[Tang 11] H. Tang, N. Joshi et A. Kapoor. Learning a blind measure of perceptual image quality. Conference on Computer Vision Pattern Recognition. IEEE, 2011.

[Tang 13] X. Tang, W. Luo et X. Wang. Content-based photo quality assessment. Transactions on Multimedia, 2013.

[Tao 15] P. Tao, J. Cao, S. Li, X. Liu et L. Liu. Mesh saliency via ranking unsalient patches in a descriptor space. Computers & Graphics, 2015.

[Tao 16] Y. Tao, Q. Wang, W. Chen, Y. Wu et H. Lin. Similarity voting based viewpoint selection for volumes. Computer Graphics Forum. Wiley Online Library, 2016.

[Tasse 15] F. P. Tasse, J. Kosinka et N. Dodgson. Cluster-based point set saliency. International Conference on Computer Vision, ICCV, 2015.

[Taubin 95] G. Taubin. Estimating the tensor of curvature of a surface from a polyhedral approximation. International Conference on Computer Vision, ICCV, 1995.

[Tofallis 14] C. Tofallis. Add or multiply ? A tutorial on ranking and choosing with multiple criteria. INFORMS Transactions on education, 2014.

[Toshev 09] A. Toshev, Aand Makadia et K. Daniilidis. Shape-based object recognition in videos using 3D synthetic object models. Conference on Computer Vision Pattern Recognition, 2009.

[Tuytelaars 04] T. Tuytelaars et L. Van Gool. Matching Widely Separated Views Based on Affine Invariant Regions. International Journal on Computer Vision, IJCV, 2004.

[Tuytelaars 06] T. Tuytelaars. Local Invariant Features : What ? Why ? When ? How ? European Conference on Computer Vision, ECCV, 2006.

[Vaswani 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser et Il Polosukhin. Attention is all you need. Advances in neural information processing systems, NIPS, 2017.

[Vázquez 01] P.-P. Vázquez, M. Feixas, M. Sbert et W. Heidrich. Viewpoint selection using viewpoint entropy. Vision Modeling and Visualization Conference. Citeseer, 2001.

[Vázquez 03] P.-P. Vázquez, M. Feixas, M. Sbert et W Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modelling. Computer Graphics Forum, 2003.

[Vázquez 09] P.-P. Vázquez. Automatic view selection through depth-based view stability analysis. The Visual Computer, 2009.

[Verdie 15] Y. Verdie, K. Yi, P. Fua et V. Lepetit. Tilde : A temporally invariant learned detector. Conference on Computer Vision Pattern Recognition, 2015.

[Vieira 09] T. Vieira, A. Bordignon, A. Peixoto, G. Tavares, H. Lopes, L. Velho et T. Lewiner. Learning good views through intelligent galleries. Computer Graphics Forum, 2009.

[Viswanatha Reddy 20] G. Viswanatha Reddy, S. Mukherjee et M. Thakur. Measuring photography aesthetics with deep CNNs. IET Image Processing, 2020.

[Wang 15] S. Wang, N. Li, S. Li, Z. Luo, Z. Su et H. Qin. Multi-scale mesh saliency based on low-rank and sparse analysis in shape feature space. Computer Aided Geometric Design, 2015.

[Wang 16] X. Wang, D. Lindlbauer, C. Lessig, M. Maertens et M. Alexa. Measuring the visual salience of 3d printed objects. Computer graphics and applications, 2016.

[Wang 17] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang et X. Tang. Residual attention network for image classification. Conference on Computer Vision Pattern Recognition, pages 3156–3164, 2017.

[Wang 18] X. Wang, S. Koch, K. Holmqvist et M. Alexa. Tracking the gaze on objects in 3D : How do people really look at the bunny ? ACM transactions on graphics, TOG, 2018.

[Wang 19] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng et J. Jia. Underexposed photo enhancement using deep illumination estimation. Conference on Computer Vision Pattern Recognition, 2019.

[Wong 09] L.-K. Wong et K.-L. Low. Saliency-enhanced image aesthetics class prediction. International Conference on Image Processing. IEEE, 2009.

[Wu 08] C. Wu, B. Clipp, X. Li, J. M. Frahm et M. Pollefeys. 3D model matching with Viewpoint-Invariant Patches (VIP). Conference on Computer Vision Pattern Recognition, 2008.

[Wu 13] J. Wu, X. Shen, W. Zhu et L. Liu. Mesh saliency with global rarity. Graphical Models, 2013.

[Xiang 14] Y. Xiang, R. Mottaghi et S. Savarese. Beyond PASCAL : A Benchmark for 3D Object Detection in the Wild. Conference on Applications of Computer Vision, WACV, 2014.

[Xiang 16] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas et S. Savarese. ObjectNet3D : A Large Scale Database for 3D Object Recognition. European Conference on Computer Vision, ECCV, 2016.

[Xu 17] C. Xu, L. Zhang, L. Cheng et R. Koch. Pose Estimation from Line Correspondences : A Complete Analysis and a Series of Solutions. Transactions on pattern analysis and machine intelligence, 2017.

[Xu 20] M. Xu, J.-X. Zhong, Y. Ren, S. Liu et G. Li. Context-aware attention network for predicting image aesthetic subjectivity. ACM International Conference on Multimedia, ACMMM, 2020.

[Xu 21] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai et Z. Liu. End-to-End Semi-Supervised Object Detection with Soft Teacher. CoRR (arXiv), 2021.

[Yang 08] W. Yang, D. Yi, Z. Lei, J. Sang et S. Z. Li. 2D–3D face matching using CCA. International Conference on Automatic Face Gesture Recognition, FG, 2008.

[Yang 19] C. Yang, Y. Li, C. Liu et X. Yuan. Deep learning-based viewpoint recommendation in volume visualization. Journal of visualization, 2019.

[Yang 21] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan et J. Gao. Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv :2107.00641, 2021.

[Yarbus 13] A. L. Yarbus. Eye movements and vision. Springer, 2013.

[Yeung 08] R. W. Yeung. Information theory and network coding. Springer Science & Business Media, 2008.

[Yi 16] K. Moo Yi, E. Trulls, V. Lepetit et P. Fua. Lift : Learned invariant feature transform. European Conference on Computer Vision, ECCV. Springer, 2016.

[Yuan 12] L. Yuan et J. Sun. Automatic exposure correction of consumer photographs. European Conference on Computer Vision, ECCV. Springer, 2012.

[Zhai 20] G. Zhai et X. Min. Perceptual image quality assessment : a survey. Science China Information Sciences, 2020.

[Zhan 19] Y. Zhan, Y. Gao et L. Y. Xie. Aesthetic feature analysis and classification of Chinese traditional painting. Journal of Beijing University of Aeronautics and Astronautics, 2019.

[Zhang 12] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu et C. Chen. Probabilistic graphlet transfer for photo cropping. Transactions on Image Processing, TIP, 2012.

[Zhang 15] L. Zhang, L. Zhang et A. Bovik. A feature-enriched completely blind image quality evaluator. Transactions on Image Processing, TIP, 2015.

[Zhang 16] F. Zhang et B. Roysam. Blind quality metric for multidistortion images based on cartoon and texture decomposition. Signal Processing Letters, 2016.

[Zhang 20] Y. Zhang, G. Fei et G. Yang. 3D viewpoint estimation based on aesthetics. IEEE Access, 2020.

[Zhang 21] J. Zhang, Y. Miao et J. Yu. A comprehensive survey on computational aesthetic evaluation of visual art images : Metrics and challenges. IEEE Access, 2021.

[Zhang 22] X. Zhang, Q. Song et G. Liu. Multimodal Image Aesthetic Prediction with Missing Modality. Mathematics, 2022.

[Zhao 14] S. Zhao, W. T. Ooi, A. Carlier, G. Morin et V. Charvillat. Bandwidth adaptation for 3D mesh preview streaming. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2014.

[Zhao 15] L. Zhao, S. Liang, J. Jia et Y. Wei. Learning best views of 3D shapes from sketch contour. The Visual Computer, 2015.

[Zhong 09] Y. Zhong. Intrinsic shape signatures : A shape descriptor for 3D object recognition. International Conference on Computer Vision, ICCV, 2009.

[Zhu 14] J.-Y. Zhu, A. Agarwala, A. A. Efros, E. Shechtman et J. Wang. Mirror mirror : Crowdsourcing better portraits. ACM transactions on graphics, TOG, 33, 2014.

[Zhu 20] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong et C. Li. Multimodal summarization with guidance of multimodal reference. Conference on Artificial Intelligence, 2020.

[Zhu 23] H. Zhu, Z. Shao, Y. Zhou, G. Wang, P. Chen et L. Li. Personalized Image Aesthetics Assessment with Attribute-guided Fine-grained Feature Representation. ACM International Conference on Multimedia, ACMMM, 2023.

[Zusne 70] L. Zusne. Visual Perception of Form. Academic Press, 1970.

Popularized abstract

In our daily lives, we encounter many 3D objects, but we often find ourselves constrained to perceive them through 2D supports. How can we choose the best view angle for these objects to maximize our understanding and identify them unambiguously ? This is the main subject of our work.

We explore two main approaches to answer this question. First, we analyze 2D images of 3D objects to determine which view offered the most relevant information. Secondly, we study directly the 3D meshes of the objects to identify the most representative and easily comprehensible views.

Our approaches are based on geometric criteria rather than aesthetic considerations. We develop two algorithms based mainly on the detection of salient points on objects in order to select the most informative viewpoints. This extraction of salient features corresponds to the essential information of the object, i.e. the minimum set of attributes sufficient to characterize and identify the object. Various techniques and features specific to each modality studied, are also exploited.

To validate our approaches, in the case of image analysis, we compare our method with traditional neural networks. As for the direct analysis of 3D meshes, we conduct studies with human participants to assess the relevance of the selected viewpoints. The results show that our methods outperform traditional approaches in terms of representativeness and ease of understanding of 3D objects.

In summary, our research is designed to help users better apprehend 3D objects on 2D supports by selecting the most relevant and representative viewpoints.

Résumé vulgarisé

Notre quotidien regorge d'objets 3D, mais nous sommes souvent contraints de les observer sur des supports 2D. Comment choisir le meilleur angle de vue pour ces objets afin de maximiser notre compréhension et pouvoir les identifier sans ambiguïté ? C'est dans ce contexte qu'interviennent nos travaux.

Nous avons exploré deux grandes approches pour répondre à cette question. Tout d'abord, nous avons analysé les images 2D des objets 3D pour déterminer quels angles de vue offraient le plus d'informations pertinentes. Ensuite, nous avons examiné directement les maillages 3D des objets pour identifier les perspectives les plus représentatives et aisément compréhensibles.

Nos approches reposent sur des critères géométriques plutôt que sur des considérations esthétiques. Nous avons développé deux algorithmes basés principalement sur la détection de points saillants des objets afin de sélectionner les points de vue les plus informatifs. Cette extraction de données saillantes correspondant à l'information essentielle de l'objet

étudié, c'est-à-dire à l'ensemble minimal d'attributs suffisant pour caractériser et identifier l'objet. Diverses techniques et caractéristiques spécifiques à chaque modalité examinée ont également été exploitées.

Pour valider nos approches, dans le cas de l'étude des images, nous avons comparé nos résultats à ceux obtenus par des réseaux de neurones classiques. Quant à l'analyse directe des maillages 3D, nous avons mené des études avec des participants humains pour évaluer la pertinence des points de vue sélectionnés. Les résultats ont montré que nos méthodes surpassaient les approches traditionnelles en termes de représentativité et de facilité de compréhension des objets 3D.

En résumé, notre recherche vise à aider les utilisateurs et utilisatrices à mieux appréhender les objets 3D sur des supports 2D en sélectionnant les angles de vue les plus pertinents et représentatifs.

Abstract

In our three-dimensional world, objects are often represented in two dimensions, mainly visualized through 2D displays such as paper catalogs or computer screens. Selecting the viewpoint from which to observe a 3D object on a 2D display has a significant impact on its recognition, the identification of its characteristics and its functional use. This task is crucial in various applications related to video games, medical imaging, architecture and industrial design.

This thesis proposes to automatically select the most relevant 2D viewpoint for a given 3D object, by quantifying the relevance of a viewpoint as the visible « essential object characteristics », i.e. the information necessary to identify the object or its main attributes. We propose a method for automatically determining the most relevant 2D viewpoint by extracting and quantifying the essential information from each available viewpoint.

Two aspects were explored in this thesis. The first evaluates the relevance of fixed viewpoints offered by textured images relative to a 3D object. The geometry of this object is defined by a 3D model. The second aspect focuses on the analysis of a geometric 3D model defined by an untextured 3D mesh. We propose to automatically determine its most representative viewpoint, the one that allows unambiguous identification and understanding.

For the first axis of research, we are developing an approach based on the detection of salient object features while filtering out irrelevant information relative to appearance, such as texture, taking advantage of both 2D images and 3D models. We introduce a relevance score, derived from geometric attributes and photography-related recommendations, which enables us to rank all the viewpoints that are available. To validate the approach, we compare the relevance scores with confidence scores obtained from learning methods, and the results obtained show the interest and effectiveness of the proposed method.

In the second part, a geometric method is proposed to determine the most representative viewpoint of an object, considering its functionnal nature rather than its aestheticism. More specifically, we propose to evaluate the amount of visible surface as well as intrinsic saliency, while weighting each visible vertex according to viewing angle. The results were validated by a user study in order to guarantee the consistency with human perception. We carried out a comparison with two state-of-the-art methods for best view selection, based on mean curvature and saliency.

Résumé

Dans notre monde tridimensionnel, les objets sont souvent représentés en deux dimensions, principalement visualisés à travers des supports 2D tels que des catalogues en papier ou des écrans d'ordinateur. La sélection du point de vue pour observer un objet en 3D sur un support 2D a un impact significatif sur son identification, la visualisation de ses caractéristiques et la compréhension de son utilité. Cette tâche est primordiale dans différents domaines applicatifs tels que les jeux vidéos, l'imagerie médicale, l'architecture ou la conception industrielle.

Cette thèse se penche sur la problématique du choix du point de vue 2D le plus adapté pour un objet 3D donné, avec pour objectif de mesurer la pertinence de ce point de vue au regard des « informations essentielles » de l'objet, c'est-à-dire les informations qui permettent de le reconnaître et d'en extraire les attributs caractéristiques. Nous proposons une méthode pour déterminer automatiquement le point de vue 2D le plus pertinent en extrayant et en quantifiant les « informations essentielles » de chaque point de vue disponible.

Deux axes de recherche ont été explorés dans cette étude. Le premier axe concerne l'évaluation de la pertinence des points de vue fixes offerts par des images texturées par rapport à un objet 3D. La géométrie de cet objet est définie par un modèle 3D. Le deuxième axe se concentre sur l'analyse d'un modèle 3D géométrique défini par un maillage 3D non texturé de l'objet. Nous proposons de déterminer automatiquement le point de vue le plus représentatif de l'objet, celui qui permet une identification et une compréhension sans ambiguïté.

Pour notre premier axe de recherche, nous développons une approche s'appuyant sur la détection des caractéristiques saillantes de l'objet tout en filtrant les informations non pertinentes comme celles liées à l'apparence, comme la texture, en tirant parti à la fois des images 2D et des modèles 3D. Nous introduisons un score de pertinence, dérivé des attributs géométriques et des recommandations liées à la photographie, qui nous permet de classer tous les points de vue disponibles. Pour valider l'approche, nous comparons les scores de pertinence avec les scores de confiance obtenus à partir de méthodes d'apprentissage, et les résultats obtenus montrent l'intérêt et l'efficacité de la méthode proposée.

Lors de la seconde partie, une méthode géométrique est proposée pour déterminer le point de vue le plus représentatif d'un objet en considérant son utilisation plutôt que son esthétisme. Plus précisément, nous proposons d'évaluer la quantité de surfaces visibles ainsi que la saillance intrinsèque, tout en pondérant chaque sommet visible

en fonction de l'angle de vue. Les résultats ont été validés par une étude utilisateur et utilisatrice, tout en assurant la cohérence avec la perception humaine.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Analyse de contenus visuels en 2D et en 3D : introduction à la notion de points de vue pertinents d'un objet 3D

Mots clés : Images 2D - Modèles 3D, Saillance Curviligne, Saillance 3D, Score de pertinence, Sélection de la meilleure vue, Etude utilisateur et utilisatrice

Résumé : Dans notre monde tridimensionnel, les objets sont souvent représentés en deux dimensions, principalement visualisés à travers des supports 2D tels que des catalogues en papier ou des écrans d'ordinateur. La sélection du point de vue pour observer un objet en 3D sur un support 2D a un impact significatif sur son identification, la visualisation de ses caractéristiques et la compréhension de son utilité. Cette tâche est primordiale dans différents domaines applicatifs tels que les jeux vidéos, l'imagerie médicale, l'architecture et la conception industrielle. Cette thèse se penche sur la problématique du choix du point de vue 2D le plus adapté pour un objet 3D donné, avec pour objectif de mesurer la pertinence de ce point de vue au regard des "informations essentielles" de l'objet, c'est-à-dire les informations qui permettent de le reconnaître et d'en extraire les attributs caractéristiques. Nous proposons une méthode pour déterminer automatiquement le point de vue 2D le plus pertinent en extrayant et en quantifiant les "informations essentielles" de chaque point de vue disponible. Deux axes de recherche ont été explorés dans cette étude. Le premier axe concerne l'évaluation de la pertinence des points de vue fixes offerts par des images texturées par rapport à un objet 3D. La géométrie de cet objet est définie par un modèle 3D. Le deuxième axe se concentre sur l'analyse d'un modèle 3D géométrique défini par un maillage 3D non texturé de l'objet. Nous proposons de déterminer automatiquement le point de vue le plus représentatif de l'objet, celui qui permet une identification et une compréhension sans ambiguïté. Pour notre premier axe de recherche, nous développons une approche s'appuyant sur la détection des caractéristiques saillantes de l'objet tout en filtrant les informations non pertinentes comme l'apparence, en tirant parti à la fois des images 2D et des modèles 3D. Nous introduisons un score de pertinence, dérivé des attributs géométriques et des recommandations liées à la photographie, qui nous permet de classer les points de vue. Pour valider l'approche, nous comparons les scores de pertinence avec les scores de confiance obtenus à partir de méthodes d'apprentissage, et les résultats obtenus montrent l'intérêt et l'efficacité de la proposition. Lors de la seconde partie, une méthode géométrique est proposée pour déterminer le point de vue le plus représentatif d'un objet en considérant son utilisation plutôt que son esthétisme. Plus précisément, nous proposons d'évaluer la quantité de surfaces visibles ainsi que la saillance intrinsèque, tout en pondérant chaque sommet visible en fonction de l'angle de vue. Les résultats ont été validés par une étude utilisateur et utilisatrice, tout en assurant la cohérence avec la perception humaine. De plus, pour démontrer l'efficacité de notre approche, nous avons réalisé une comparaison avec deux méthodes de l'état de l'art qui proposent également des techniques de sélection de la meilleure vue d'un objet. Cependant, contrairement à ces deux méthodes, notre approche pour sélectionner le meilleur point de vue est plus élaborée, impliquant une combinaison d'attributs géométriques et sémantiques propres à l'objet 3D étudié. En conclusion, nos contributions portent sur la sélection automatique du point de vue le plus pertinent pour un objet 3D donné. Autrement dit, nous identifions la vue optimale qui représente le mieux l'objet 3D, c'est-à-dire, celle qui permet de visualiser le plus de caractéristiques essentielles de l'objet 3D, facilitant ainsi son identification et sa compréhension. Dans nos deux axes de travail, cette identification se réalise à l'aide d'une quantification de la pertinence, basée sur l'extraction de l'information essentielle de l'objet étudié.

Title: Analysis of 2D and 3D visual content: introduction to the notion of relevant viewpoints of a 3D object

Key words: 2D Images - 3D Models, Curvilinear saliency, 3D saliency, Relevance score, Best view selection, User study

Abstract: In our three-dimensional world, objects are often represented in two dimensions, mainly visualized through 2D displays such as paper catalogs or computer screens. Selecting the viewpoint from which to observe a 3D object on a 2D display has a significant impact on its identification, the identification of its characteristics and the understanding of its purpose. This task is crucial in various application domains such as video games, medical imaging, architecture and industrial design. This thesis proposes to automatically select the most relevant 2D viewpoint for a given 3D object, by quantifying the relevance of a viewpoint as the visible "essential object characteristics", i.e. the information necessary to identify the object or its important attributes. We propose a method for automatically determining the most relevant 2D viewpoint by extracting and quantifying the essential information from each available viewpoint. Two lines of research were explored in this study. The first evaluates the relevance of fixed viewpoints offered by textured images relative to a 3D object. The geometry of this object is defined by a 3D model. The second axis focuses on the analysis of a geometric 3D model defined by an untextured 3D mesh representing the object. We propose to automatically determine its most representative viewpoint, the one that allows unambiguous identification and understanding. For our first axis of research, we are developing an approach based on the detection of salient object features while filtering out irrelevant information such as appearance, taking advantage of both 2D images and 3D models. We introduce a relevance score, derived from geometric attributes and photography-related recommendations, which enables us to rank viewpoints. To validate the approach, we compare the relevance scores with confidence scores obtained from learning methods, and the results obtained show the interest and effectiveness of the proposal. In the second part, a geometric method is proposed to determine the most representative viewpoint of an object, considering its purpose rather than its aestheticism. More specifically, we propose to evaluate the amount of visible surface as well as intrinsic saliency, while weighting each visible vertex according to viewing angle. The results were validated by a user study, while ensuring consistency with human perception. In addition, to demonstrate the effectiveness of our approach, we carried out a comparison with two state-of-the-art methods that also offer techniques for selecting the best view of an object. However, unlike those methods, our approach to selecting the best viewpoint is more sophisticated, involving a combination of geometric and semantic attributes inherent to the studied 3D object. In conclusion, our contributions focus on the automatic selection of the most relevant viewpoint for a given 3D object. In other words, we identify the optimal view that accurately and exhaustively represents the object's essential features, making it easier to identify and understand. In our two axes of work, this identification is achieved through a quantification of relevance, based on the extraction of essential information from the object under study.