Towards an analysis and comprehension of image quality

Quyet Tien Le

Université Grenoble Alpes

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **SIGNAL IMAGE PAROLE TELECOMS**

Arrêté ministériel : 25 mai 2016

Présentée par
Quyet-Tien LE

Thèse dirigée par **Alice CAPLIER, Professeur, Université Grenoble Alpes** et
encadrée par **Patricia LADRET, Université Grenoble Alpes** et
encadré par **Huu-Tuan NGUYEN, Vietnam Maritime University**

préparée au sein du
Grenoble, Images, Parole, Signal du Automatiqe (GIPSA)
dans l'École Doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)

Vers une analyse et une compréhension de la qualité d'image

Towards an analysis and comprehension of image quality

Thèse soutenue publiquement le **16 novembre 2020**,
devant le jury composé de :

Mme. Fan YANG
Professeur, Universite De Bourgogne, Rapportrice
M. Patrick LE CALLET
Professeur, Universite De Nantes, Rapporteur
M. Denis PELLERIN
Professeur, Universite Grenoble Alpes, Président
M. Arnaud LIENHARD
Docteur-ingenieur, Neovision Grenoble, Examinateur
Mme. Alice CAPLIER
Professeur, Universite Grenoble Alpes, Directrice de thèse

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE EEATS

Electronique, Electrotechnique, Automatique, Traitement du Signal

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble Alpes

Mention : Signal image parole telecoms

Présentée et soutenue par

Quyet-Tien LE

Towards an analysis and comprehension of image quality

Thèse dirigée par Alice CAPLIER

Grenoble Image Parole Signal Automatique (GIPSA)

16 novembre 2020

Jury :

Rapporteurs :	Mme. Fan YANG	-	Laboratoire LE2I
	M. Patrick LE CALLET	-	Laboratoire LS2N
Directrice :	Mme. Alice CAPLIER	-	Laboratoire GIPSA-lab
Co-encadrants :	Mme. Patricia LADRET	-	Laboratoire GIPSA-lab
	M. NGUYEN Huu Tuan	-	Vietnam Maritime University
Président :	M. Denis PELLERIN	-	Laboratoire GIPSA-lab
Examinateur :	M. Arnaud LIENHARD	-	Neovision

Acknowledgments

PhD life is a long journey that no one can finish it alone. It is lucky to me that my family, my supervisors and my friends always support and encourage me in my journey.

Firstly, I would like to express my deep gratitude to my supervisors in France, Prof. Alice Caplier and Prof. Patricia Ladret for all of their help from the beginning days of my journey not only in work but also in life. They taught me everything about research and helped me through tough periods in life.

I would like to thank Dr. Nguyen Huu Tuan, my supervisor in Vietnam. He is my supervisor not only during PhD period but also during my bachelor and master periods. He has worked with me and helped me for over ten years.

A great thanks to Prof. Denis Pellerin for accepting to become the president of my thesis committee. Many thanks to Prof. Le Callet and Prof. Yang for their acceptances to be the reviewers for my thesis manuscript. I would like to thank Dr. Lienhard for agreeing to become the examiner for my thesis. Dr. Lienhard was also a member of my CSI committee during my PhD period. I appreciate their time reading this thesis and based on their comments, this thesis has been improved.

I wish to thank my friends in GIPSA Lab : 2 Pedro*s*, my officemates, Jitu, Karina, Luisa, Saloua, Dora, Fateme, Julien, Ivan, Dawood, Maria, Ludo, Bruce, Phuc, Tan, Hung, Phong for their help and friendship. They make GIPSA my second home.

A special thanks to my Vietnamese friends at Ile Verte residence and in Grenoble : Le Minh Thong, Le Van Thao, Pham Van Hung, Nguyen Trung Hieu's family, Truong Ba Luu's family, Pham Hoang Lam's family, Tran Nguyen Viet Khoa and Vo Trong Hong. I will never forget the period when I was surrounded and encouraged by all of them.

I am grateful to acknowledge that my PhD study was funded by the Vietnamese government scholarship program 911.

Finally, I would like to express my greatest gratitude to my family, my wife, my children, my parents, my sister. Without their love, support and sacrifice, this thesis could never be done.

Table des matières

Table des figures

Liste des tableaux

Table of acronyms

AUC	Area Under Curve
CIAA	Close-up Image Aesthetic Assessment
CNN	Convolutional Neural Network
DOF	Depth Of Field
GIAA	General Image Aesthetic Assessment
HDR	High Dynamic Range
HVS	Human Visual System
IAA	Image Aesthetic Assessment
INA	Image Naturalness Assessment
IQA	Image Quality Assessment
LCIC	Large field Close-up Image Classification
LIAA	Large field Image Aesthetic Assessment
RB	Region Of Interest and Background
ROI	Region Of Interest
ROIE	Region Of Interest Extraction
SDR	Standard Dynamic Range
TMI	Tone Mapped Image
TMO	Tone Mapping Operator

Résumé

Le système visuel humain (HVS) [Gao+10] ; [Mar10] ; [WT97] joue un rôle important dans la vie humaine. Le HVS aide les humains à indiquer leur orientation, à détecter, identifier et reconnaître des objets et à effectuer de nombreuses tâches quotidiennes. Le système comprend 2 composantes principales : un organe sensoriel qui collecte les informations visuelles et une partie du système neuronal qui traite les informations visuelles collectées. L'organe sensoriel et le système neuronal du HVS sont respectivement liés à l'acuité visuelle humaine et à la perception visuelle humaine. Bien qu'il existe certaines relations entre les deux concepts, l'acuité visuelle humaine et la perception visuelle humaine sont complètement différentes. L'acuité visuelle humaine fait référence à la clarté de la vision qui dépend de la netteté du foyer rétinien de l'œil, tandis que la perception visuelle humaine est la capacité de déchiffrer, d'analyser les informations visuelles reçues par les yeux humains du milieu environnant. En fait, une personne peut avoir des problèmes de perception visuelle même si elle a une acuité visuelle normale. La réception et le traitement des informations visuelles sont toujours des tâches humaines essentielles. Depuis des millions d'années, l'homme a essayé de capturer des informations visuelles : de simples symboles de peintures rupestres anciennes aux photos satellites numériques, des images en niveaux de gris aux images à haute gamme dynamique. Ce sont des efforts humains pour capturer et analyser les informations visuelles du monde [Kra+17].

Ces dernières années, avec le développement de la technologie, les appareils numériques sont omniprésents dans la vie moderne. Il existe de plus en plus de dispositifs intelligents intégrés à des appareils photo numériques tels que les smartphones, les tablettes, les ordinateurs portables, les smartwatches. Par rapport au passé, où une personne devait se rendre dans un studio pour prendre une photo et devait attendre des heures pour recevoir la photo, les dispositifs numériques nous ont permis de capturer et de stocker plus facilement des informations visuelles afin de pouvoir prendre une photo à tout moment et en tout lieu. Le nombre d'images a donc augmenté de façon spectaculaire de jour en jour et le stockage des utilisateurs peut être rempli très rapidement. Il est donc nécessaire d'évaluer les photos pour conserver les meilleures et supprimer les plus mauvaises. La sélection manuelle prend beaucoup de temps et est assez complexe, de sorte que l'évaluation automatique de la qualité des images pourrait permettre d'accélérer la tâche. La tâche d'évaluation est effectuée sur la base de la perception visuelle humaine. La Fig. 1 montre une vue d'ensemble des facteurs d'image affectant la perception visuelle humaine, classés en 2 groupes : le contenu de l'image et la qualité de l'image.

Le contenu des images a une grande influence sur la perception visuelle de l'homme. En ce qui concerne le contenu des images, il y a 3 facteurs : Le message à l'intérieur, l'inspiration émotionnelle et l'originalité de l'image. En regardant les exemples de la Fig. 1, la première photo est un exemple de "message intérieur". Sur cette photo, on voit un oiseau coincé dans un sac en plastique. Bien que le contenu semble simple, il pourrait contenir un message caché lié à l'environnement comme "Sauvons les animaux", "Arrêtons de consommer des sacs en plastique" ou "Notre planète est détruite". En ce qui concerne la deuxième photo, certaines personnes n'ont peut-être pas de sentiments particuliers à son égard, mais l'accolade entre la

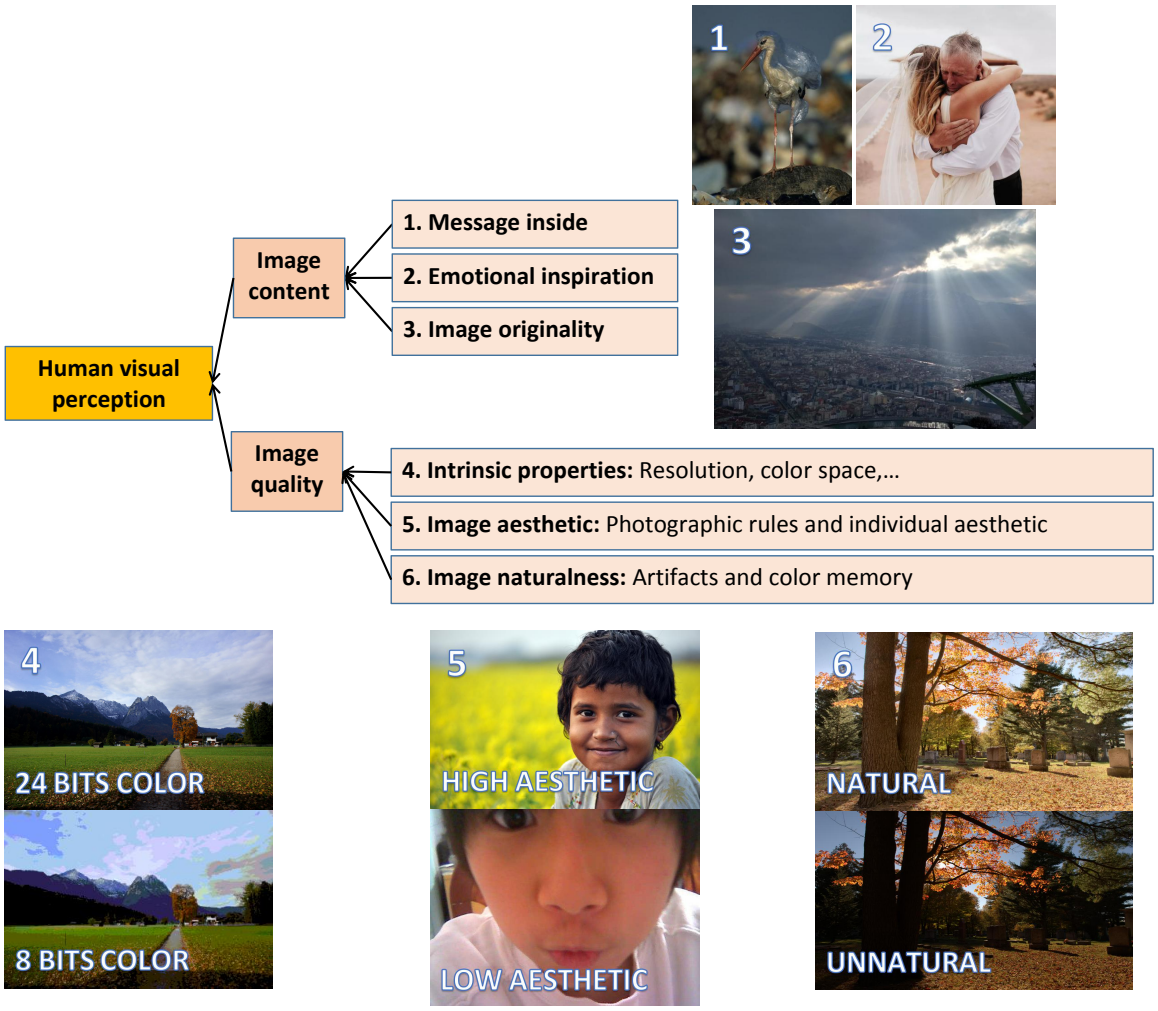


FIGURE 1 – Aperçu des aspects de l'image ayant une influence sur la perception visuelle humaine.

mariée et son père pourrait rappeler aux autres personnes les membres de leur famille ou un souvenir personnel. La valeur de la photo est "l'inspiration émotionnelle". Dans la troisième photo, un moment rare de ciel nuageux avec des rayons lumineux rend la photo différente des autres photos. Bien qu'il n'y ait pas de message caché ou d'inspiration émotionnelle dans ce cas, l'originalité rend la photo spéciale. En général, les 3 facteurs sont exprimés par le contenu de l'image plutôt que par son apparence. Bien qu'il y ait de plus en plus de personnes ayant des connaissances photographiques (sur les appareils : objectif, puce de l'appareil photo, sur les propriétés de la photo : éclairage, résolution, netteté, sur les réglages : ouverture, temps d'exposition, longueur focale, sur les règles photographiques : composition, contraste, profondeur de champ), le contenu de l'image est surtout exploité par des photographes professionnels. Pour l'évaluation des images pour les utilisateurs réguliers, la qualité de l'image est le principal critère.

Dans notre étude, la qualité de l'image n'est pas censée être liée au contenu de l'image. La notion de qualité peut être abordée sous différents angles. Dans le domaine des télécommunications, la qualité de service est définie comme "l'ensemble des caractéristiques d'un service de télécommunications qui influent sur sa capacité à satisfaire les besoins déclarés et implicites de l'utilisateur du service". En ce qui concerne cette définition de la qualité, la qualité de l'image est considérée comme un problème technique puisqu'elle est définie sur la base de propriétés intrinsèques uniquement : résolution, espace colorimétrique, profondeur de couleur, format de l'image,... (voir Fig. 1(4) et Fig. 2) Cette définition fait principalement référence aux propriétés de l'image elle-même et ne mentionne aucun facteur induit par les spectateurs. Dans le passé, les propriétés intrinsèques étaient le facteur principal lorsqu'il s'agissait de la qualité de l'image, car les différences techniques telles que la résolution, la profondeur de couleur, les points étaient importantes. Cependant, les appareils numériques intégrés aux caméras sont de plus en plus populaires en raison de leur prix abordable. Il est donc plus facile pour les utilisateurs de posséder une caméra numérique capable de produire des photos à haute résolution et à gamme dynamique standard. Ainsi, les différences de propriétés intrinsèques ont été réduites et le rôle des propriétés intrinsèques dans la qualité de l'image est devenu insignifiant.

Un autre concept de qualité axé sur les facteurs liés aux utilisateurs est la qualité de l'expérience qui est définie comme "le degré de plaisir ou d'ennui de l'utilisateur d'une application ou d'un service". Elle résulte de la satisfaction de ses attentes en ce qui concerne l'utilité et/ou la jouissance de l'application ou du service à la lumière de la personnalité de l'utilisateur et de son état actuel" " [Bru+13]. Il apparaît que l'évaluation de la qualité considérée par différents téléspectateurs dans des conditions différentes ne pourrait pas être la même. Selon la seconde approche, la qualité de l'image est définie comme l'appréciation d'un observateur sur une photo qui repose sur 2 notions : l'esthétique de l'image et le naturel de l'image. Plus précisément, l'esthétique de l'image est la mesure de la façon dont une photo répond esthétiquement à l'attente de l'observateur (voir Fig. 1(5) et Fig. 3) tandis que la définition du naturel de l'image est à la fois liée aux artefacts induits par certains algorithmes de traitement d'image et au sentiment individuel sur la façon dont une photo correspond à la mémoire d'image [LE+20] (voir Fig. 1(6) et Fig. 5).

FIGURE 2 – Exemples de propriétés intrinsèques. La première photo a une résolution plus élevée que les autres, tandis que la profondeur de couleur de la troisième photo est plus faible que celle des deux premières.

En fait, le sujet principal de la thèse est la qualité de l'image. En se concentrant sur les 2 facteurs : l'esthétique de l'image et le naturel de l'image, nous n'allons pas considérer ni le contenu de l'image ni les propriétés intrinsèques de l'image.

La qualité de l'image est le sujet principal de cette thèse. Il y a 3 aspects principaux se référant à la qualité de l'image : les propriétés intrinsèques, l'esthétique de l'image et le naturel de l'image (cf. Fig. 1). Comme présenté précédemment, les propriétés intrinsèques reflétant les caractéristiques techniques des images ont été la première préoccupation historique concernant la qualité des images mais leur influence sur la qualité des images a été insignifiante en raison du développement de la technologie. En revanche, la qualité de l'image dépend des attentes humaines et les spectateurs exigent une photo qui soit non seulement fidèle mais aussi belle. Avec le temps, l'esthétique et le naturel de l'image sont devenus les deux facteurs les plus importants définissant la qualité de l'image. Bien qu'ils aient été identifiés comme des sujets de recherche depuis plus de trois décennies, ils constituent toujours des défis car l'esthétique et le naturel de l'image sont liés non seulement à la validité universelle, qui fait référence à la justesse ou aux affirmations sur la compréhension des spectateurs, mais aussi à la subjectivité, qui repose sur un sentiment individuel de plaisir ou de mécontentement. Les 2 principes contradictoires des 2 facteurs définissant la qualité de l'image renvoient à un débat plus général : "L'AQI est-elle subjective ou universelle ? L'esthétique de l'image est une notion abstraite qui traite de ce que ressentent les observateurs à propos de l'apparence d'une image. Cette notion est liée à ce qui se passe dans l'esprit de l'observateur lorsqu'il regarde une photo (voir les exemples dans la Fig. 3). Les questions relatives à la manière dont une photo est prise ainsi qu'à la manière dont un spectateur apprécie et critique la photo conduisent à la formation de la photographie. La photographie est l'art de créer des images durables en capturant la lumière. Bien que les opinions esthétiques ne soient pas les mêmes pour chaque spectateur, des règles de photographie ont été introduites sur la base d'aspects esthétiques descriptifs afin d'aider à prendre une meilleure photo. Cependant, tous les aspects esthétiques ne sont pas descriptibles, ce qui ne signifie pas que le respect des règles de la photographie produit toujours une photo hautement esthétique et qu'au contraire, une belle photo pourrait ne pas respecter ces règles. En photographie, il existe des règles globales se référant aux aspects esthétiques universels pour tous les types d'images et des règles particulières formées pour une catégorie d'images particulière. Certaines caractéristiques esthétiques sont bonnes pour les photos d'une catégorie spécifique et pas pour d'autres catégories d'images. En particulier, il existe deux catégories d'images ayant des règles de photographie opposées : les images à grand champ et les images en gros plan. En regardant la Fig. 4, les photos de la première rangée sont des images à grand champ (images d'une scène à grand champ prises avec une longue distance entre l'appareil photo et la scène) tandis que la deuxième rangée contient des images en gros plan (images se concentrant sur des objets en gros plan prises avec une courte distance entre l'appareil photo et les objets). Du côté des images grand champ, les photographes suivent souvent la règle du paysage pour prendre des photos grand champ. Selon la règle du paysage, une photo est divisée en 3 régions (voir exemples dans la Fig. 4) et les compositions d'une photo grand champ sont généralement disposées dans ces régions. Sur la première et la troisième photo, la ligne d'horizon est utilisée pour séparer les régions terrestres et le ciel, tandis que sur la deuxième photo, le sol du Colisée correspond à la ligne de la règle du paysage. En général, les photographes exploitent les lignes dans les scènes pour séparer les

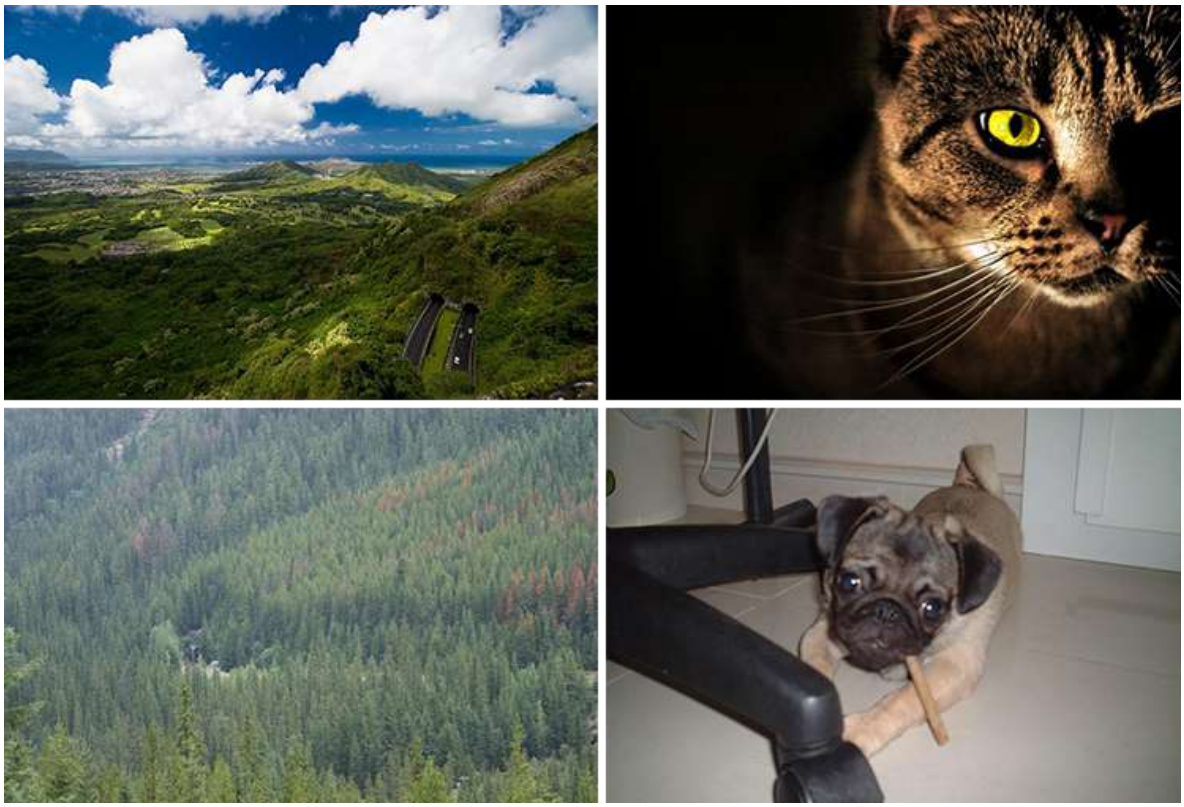


FIGURE 3 – Image aesthetic illustrations. Des exemples d'images à haute valeur esthétique figurent sur la première ligne, tandis que la deuxième ligne contient des exemples d'images à faible valeur esthétique.

régions. Une autre technique appliquée pour les images à grand champ consiste à exploiter les motifs. Par exemple, les motifs de l'herbe, des entrées et des bâtiments sont utilisés dans les Fig. 4(a), (b) et (c) respectivement. En ce qui concerne les réglages de l'appareil photo, les images grand champ sont prises avec une longue distance entre l'appareil photo et la scène, de sorte que les photographes choisissent souvent un réglage d'ouverture réduit pour obtenir une profondeur de champ importante, tandis qu'une courte distance focale est choisie pour prendre une scène large.

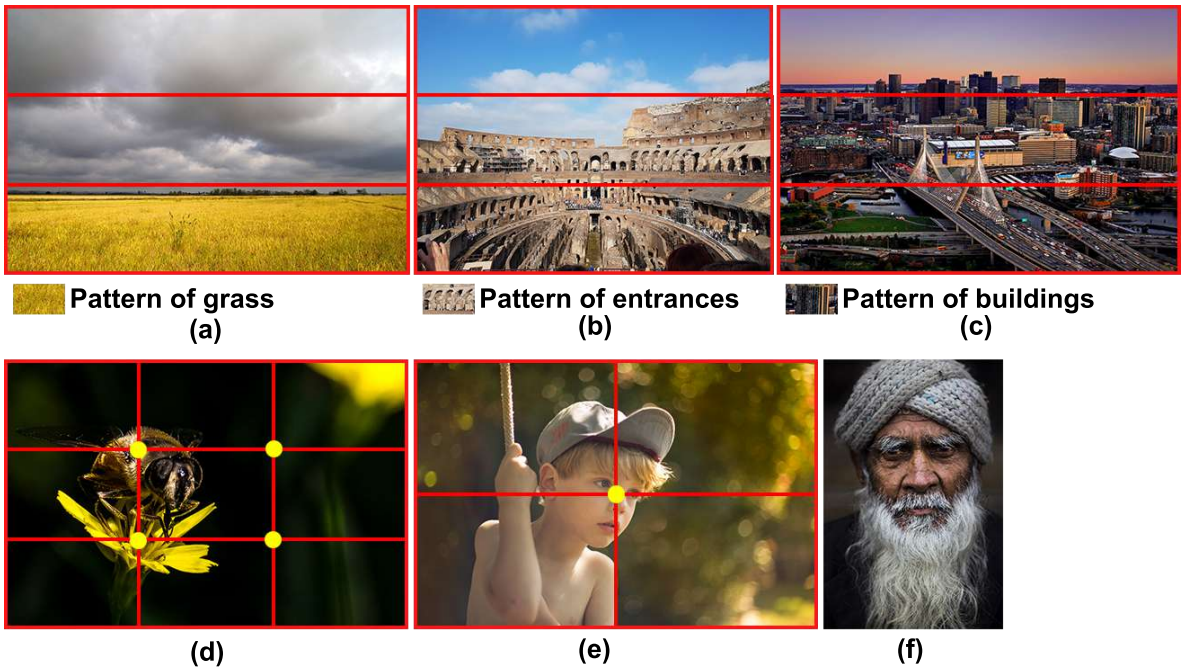


FIGURE 4 – Exemples d'images en grand champ et en gros plan.

De l'autre côté, la règle des tiers et la règle de la "mise au point centrale" sont utilisées lors de la prise d'images en gros plan (voir Fig. 4(d) et (e) respectivement). Alors que les règles pour les images à grand champ se concentrent sur les régions et les lignes, les règles pour les images en gros plan se concentrent sur les points. Selon la règle des tiers et la règle de la "mise au point centrale", les observateurs prêtent davantage attention à certains points particuliers (points jaunes dans la Fig. 4(d) et (e)), de sorte que le contenu principal des photos doit être placé à ces points (ou à proximité). Les photographes utilisent souvent l'effet "faible profondeur de champ" pour rendre l'arrière-plan flou afin de mettre en évidence le contenu principal d'une image en gros plan, de sorte que le rôle de l'arrière-plan dans l'esthétique de l'image en gros plan est insignifiant. Ainsi, une autre règle appliquée pour les photos en gros plan est "remplir le cadre". En regardant la Fig. 4(f), presque tout le cadre de la photo est recouvert du portrait du vieil homme et les zones de l'arrière-plan de cette photo sont plus petites que celles des deux premières photos en gros plan. Cependant, cette règle est rarement appliquée lors de la prise de photos à grand champ. Elle est opposée aux réglages de l'appareil photo pour la prise de vue en grand champ. Un réglage d'ouverture élevé et une

longue distance focale sont souvent réglés afin de faire la mise au point sur les objets en gros plan et d'obtenir une faible profondeur de champ lors de la prise d'une photo en gros plan.

L'évaluation esthétique des images (IAA) est basée sur des caractéristiques esthétiques et ces caractéristiques sont en quelque sorte similaires aux règles de la photographie. Il existe donc des caractéristiques universelles pouvant s'appliquer à toutes les images, mais il y a aussi des caractéristiques particulières pour une catégorie d'images spécifique qui ne pourraient pas s'appliquer aux autres catégories d'images. Il existe donc une hypothèse implicite selon laquelle l'évaluation de l'esthétique des images selon différentes vues, des critères pour chaque catégorie d'images et de la classification des images peut améliorer les performances de l'IAA.

En outre, certaines régions d'une image sont plus nettes que les autres et certaines régions sont saillantes en raison de leurs couleurs contrastées. Ces régions semblent attirer davantage l'attention des spectateurs que les autres. En regardant la Fig. 4, d'une part, les images en gros plan présentent des régions d'objet nettes (l'abeille et la fleur sur la première photo, le garçon et le vieil homme sur la deuxième et la troisième respectivement) et il semble que le contraste de couleurs entre l'arrière-plan et le premier plan soit significatif. D'autre part, le champ d'herbe, le Colisée et les scènes de ville constituent le contenu principal des 3 images grand champ et les spectateurs pourraient être davantage attirés par ces régions. Bien que les régions du ciel soient moins détaillées et ne constituent pas le contenu principal de ces photos, leur influence sur la qualité esthétique des images à grand champ est significative, de sorte que le rôle de ces régions n'est pas le même que celui de l'arrière-plan dans les images en gros plan. On suppose que la qualité esthétique d'une image est plus liée à la qualité esthétique de ces régions qu'à la qualité esthétique de l'image dans son ensemble. Cette relation peut ne pas être la même pour toutes les images et elle dépend de la catégorie de la photo considérée.

En résumé, l'esthétique de l'image est l'un des 2 principaux facteurs définissant la qualité de l'image et c'est actuellement la préoccupation la plus importante concernant la qualité de l'image, c'est pourquoi l'esthétique de l'image va être étudiée dans cette thèse. En outre, il semble que les manières de prendre, de composer ou de regarder une image esthétique d'une scène de grand champ et d'une scène de champ proche soient très différentes. C'est pourquoi, dans notre travail, nous allons considérer les deux catégories d'images pour les besoins de l'IAA. En outre, les régions aux couleurs vives et/ou très contrastées attirent davantage l'attention des spectateurs que les autres régions, de sorte que l'esthétique de l'image peut avoir une relation plus forte avec les régions saillantes que l'ensemble du cadre de la photo. Ainsi, le rôle de la segmentation des régions va être étudié pour l'IAA.

Le naturel de l'image est une notion difficile à définir. Elle est à la fois liée aux artefacts et à la correspondance entre l'apparence de l'image et la connaissance de la réalité stockée dans la mémoire des spectateurs (voir les exemples de la Fig. 5). D'une part, le caractère naturel de l'image est affecté par de forts indices non naturels détectés par les yeux des spectateurs, de sorte que la sensation de non-nature provient d'artefacts gênants induits par les capteurs de l'appareil photo, les algorithmes de traitement de l'image (compression, tone mapping, points), le format de l'image, le transfert de fichiers, etc. Ces signes peuvent aller d'imperceptibles à évidents et une photo présentant des artefacts gênants est considérée comme une photo non naturelle (voir la photo en bas à droite dans la Fig. 5).

FIGURE 5 – NNNature / illustrations contre-nature. Des exemples d'images naturelles se trouvent dans la première ligne, tandis que la deuxième ligne contient des exemples d'images non naturelles.

D'autre part, le sentiment de naturel et de contre-nature provient de l'expérience et de la mémoire du spectateur (voir la photo en bas à gauche dans la Fig. 5). Lorsqu'ils regardent une photo, les observateurs comparent la scène de la photo à la réalité extraite de leur mémoire (ce qu'ils ont vu) pour trouver des différences et des similitudes, les sentiments dépendent donc de facteurs individuels. Par exemple, en regardant la Fig. 6, il apparaît que les couleurs des feuilles en automne sont diverses et que les couleurs changent dans des conditions de luminosité différentes. Une couleur peut être familière à certains téléspectateurs mais ne pas être fidèle aux autres. Dans ce cas, certains téléspectateurs considèrent des couleurs qu'ils n'ont jamais vues comme non naturelles parce qu'ils ont déjà fixé une gamme de couleurs pour les feuilles d'automne dans leur esprit.

FIGURE 6 – Exemples de différentes couleurs de feuilles en automne.

Bien que l'esthétique de l'image soit la préoccupation la plus importante en matière de qualité de l'image, il ne suffit pas de définir précisément la qualité de l'image. Une photo à haute esthétique peut ne pas être une photo de haute qualité si elle n'est pas fidèle aux spectateurs. Le naturel de l'image est la contrepartie de l'esthétique de l'image pour définir précisément la qualité de l'image. C'est pourquoi la naturalité de l'image est étudiée dans la thèse.

Objectifs : dans cette thèse, nous allons nous concentrer sur l'esthétique et la naturalité de l'image qui sont deux facteurs impliqués dans la qualité de l'image. Les propriétés intrinsèques de l'image ne seront pas prises en compte (cf. Fig. 1).

Plus précisément, le premier aspect (l'esthétique de l'image) sera étudié pour deux catégories spécifiques d'images qui sont les images à grand champ et les images à champ proche. En partant de l'hypothèse que les catégories d'images et par conséquent la segmentation des régions (zones saillantes et arrière-plan) pourraient avoir une influence sur l'esthétique des images, notre première contribution est d'étudier l'esthétique des images évaluée en particulier sur les images avec et sans prétraitement (classification des images et segmentation des régions). Bien qu'il existe de nombreuses recherches sur l'évaluation esthétique des images (IAA), elles se concentrent principalement sur le développement de méthodes de jugement esthétique des images, alors que la question "quelle est l'influence des opérations de prétraitement dans l'IAA" n'a pas trouvé de réponse. Dans cette étude, le rôle de la classification des images et de la segmentation des régions dans l'IAA est d'abord examiné. Sur la base du rôle estimé, notre deuxième contribution est de présenter un nouveau modèle IAA impliquant la classification des images et la segmentation des régions (voir Fig. 7).

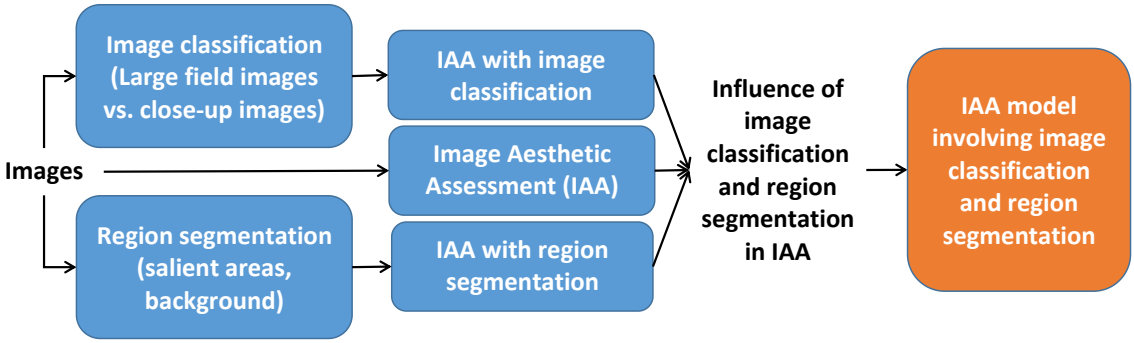


FIGURE 7 – Overview of the image aesthetic study.

Le second objectif est d'étudier le naturel des images de la carte de ton. Il y a peu de recherches sur ce sujet et il y a donc beaucoup de questions auxquelles il faut répondre. Notre contribution à ce domaine de recherche est double : proposer des expériences subjectives et développer des mesures objectives pour l'évaluation du caractère naturel des images (INA) (voir Fig. 8).

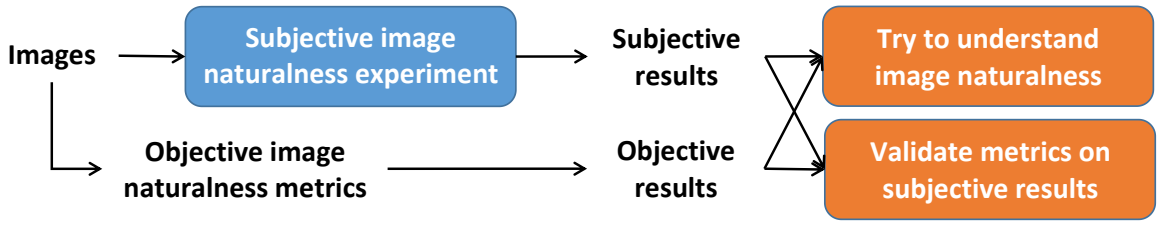


FIGURE 8 – Overview of the image naturalness study.

Enfin, les relations et les corrélations entre l'esthétique de l'image et le naturel de l'image sont étudiées.

Outline : Le contenu principal de la thèse est organisé en 4 chapitres.

Chapitre 1 décrit les deux méthodes de prétraitement proposées pour l'IAA, y compris les méthodes d'extraction de régions d'intérêt (ROIE) et les méthodes de classification d'images à grand champ/en gros plan (LCIC). La région d'intérêt (ROI) dans cette thèse est définie sur la base des informations de netteté et de couleur. Un algorithme ROIE artisanal basé sur les caractéristiques et des méthodes ROIE basées sur l'apprentissage profond sont évalués et comparés. Dans une deuxième partie, différents types de caractéristiques, y compris les caractéristiques EXIF (Exchangeable Image File Format), les caractéristiques artisanales et les caractéristiques apprises sont comparées pour répondre à la tâche du LCIC et fournir un algorithme de classification efficace.

Chapitre 2 présente l'étude de l'IAA basée sur la classification des images et la segmen-

tation des régions. En premier lieu, les performances de l'IAA avec et sans LCIC et ROIE sont considérées pour évaluer le rôle de la classification d'images et de la segmentation de régions dans l'IAA. Ensuite, en fonction de l'influence estimée de la classification des images et de la segmentation des régions, un modèle d'AIA basé sur la CILC et le ROIE est proposé. Les caractéristiques artisanales et les caractéristiques apprises sont exploitées et validées sur des données subjectives pour évaluer l'efficacité des deux types de caractéristiques dans la tâche IAA.

Chapitre 3 décrit une expérience subjective organisée en laboratoire pour recueillir les évaluations humaines sur le naturel des images cartographiées en tons. Sur la base des données subjectives obtenues lors de l'expérience, différentes mesures objectives sont validées pour la tâche INA. En outre, la définition de la naturalité des images est clarifiée et les facteurs affectant la naturalité des images sont expliqués et discutés. En outre, les relations et les corrélations entre l'esthétique et le naturel des images sont abordées.

Chapitre 4 résume les contributions de la thèse. En outre, des perspectives sur les travaux futurs sont établies.

Chapitre 1

Introduction

Sommaire

Human Visual System (HVS) [Gao+10] ; [Mar10] ; [WT97] plays an important role in human life. HVS helps humans indicating orientation, detecting, identifying and recognizing objects and performing many daily tasks. The system includes 2 main components : a sensory organ collecting visual information and a part of the neural system processing the collected visual information. The sensory organ and the neural system of HVS are related to human visual acuity and human visual perception respectively. Although there are some relations between the 2 concepts, human visual acuity and human visual perception are completely different. Human visual acuity refers to the clearness of vision depending on the sharpness of the retinal focus within the eye while human visual perception is the ability to decipher, analyze visual information received by human eyes from surrounding environment. In fact, a person could have problems with visual perception even though he/she has normal visual acuity. Receiving and processing visual information are always essential human tasks. Since millions years ago, humans have tried to capture visual information : from simple ancient cave painting symbols to digital satellite photos, from images in gray levels to High Dynamic Range images. They are human efforts to capture and analyze visual information of the world [Kra+17].

In recent years, with the development of technology, digital devices are omnipresent in modern life. There are more and more smart devices integrated with digital cameras such as smartphones, tablets, laptops, smartwatches... Comparing to the past when a person had to go to a studio to take a photo and he/she needed to wait for hours to receive the photo, digital devices have helped us capturing and storing visual information more easily so a picture could be taken whenever and wherever. It leads to the fact that the number of images has increased dramatically day by day and users' storage can be filled very fast. It leads to a need of evaluating photos to keep the best ones and to remove the worst ones. Doing the selection task by hand takes a lot of time and it is quite complex so assessing image quality automatically could help performing the task faster. The assessment task is performed based on human visual perception. Fig. 1.1 shows an overview of image factors affecting human visual perception categorized into 2 groups : image content and image quality.

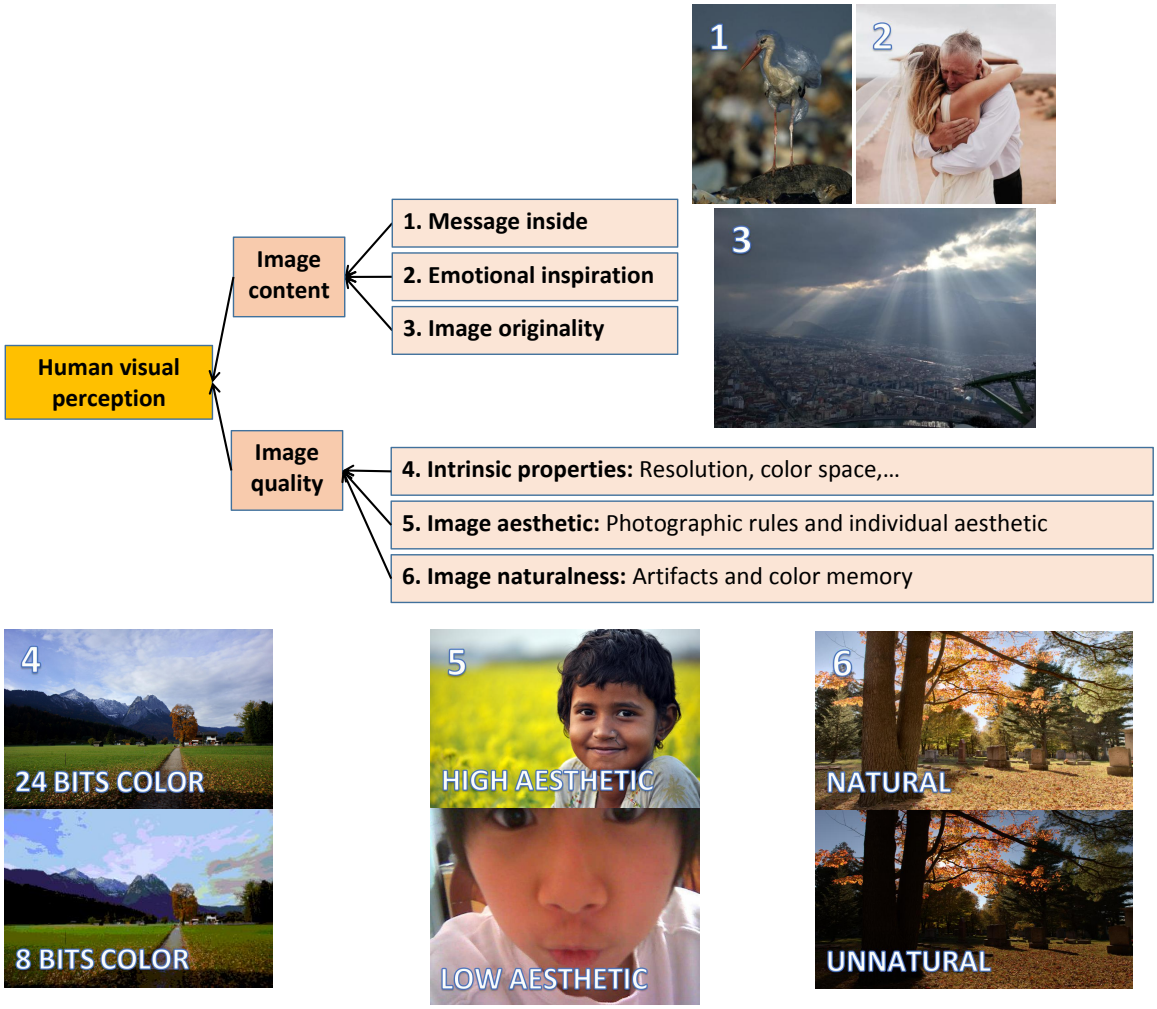


FIGURE 1.1 – Overview of image aspects having influence on human visual perception.

Image content has a great influence on human visual perception. With regard to image content, there are 3 factors : "message inside", "emotional inspiration" and "image originality". Looking at examples in Fig. 1.1, the first photo is an example for "message inside". In this photo, there is a bird stuck in a plastic bag. Although the content looks simple, it might contain a hidden message related to environment like "Let's save animals", "Stop consuming plastic bags" or "Our planet is destroyed"... Regarding the second photo, some people might not have any special feelings about it but the hug between the bride and her father could remind other people of their family members or a personal memory. The value of the photo is "emotional inspiration". In the third photo, a rare moment of a cloudy sky with light rays makes the photo different from other photos. Although there is no hidden message or emotional inspiration in this case, the originality makes the photo special. In general, the 3 factors are expressed by image content instead of image appearance. Although there are more and more people having photographic knowledge (about devices : lens, camera chip,... about photo properties : illumination, resolution, sharpness,... about settings : aperture, exposure time, focal length... about photographic rules : composition, contrast, depth of field,...), image content is mostly exploited by professional photographers. For the purpose of evaluating images for regular users, image quality is the main criterion.

In our study, image quality is not supposed to be related to image content. The concept of quality can be approached from different angles. From the approach of telecommunication area, quality of service is defined as "the totality of characteristics of a telecommunication service that bears on its ability to satisfy stated and implied needs of the user of the service" [Uni08]. With regard to this definition of quality, image quality is considered as a technical problem since it is defined based on intrinsic properties only : resolution, color space, color depth, image format,... (see Fig. 1.1(4) and Fig. 1.2) This definition mostly refers to the properties of the image itself and it does not mention any factors induced by viewers. In the past, intrinsic properties were the main factor when referring to image quality since the technical differences such as resolution, color depth,... were significant. However, digital devices integrated with cameras are more and more popular because of their affordable prices so it is easier for users to own a digital camera that is able to produce high resolution and standard dynamic range photos. Thus the differences in intrinsic properties have been reduced and the role of intrinsic properties in image quality has become insignificant.

Another concept of quality focusing on factors related to users is quality of experience that is defined as "the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [Bru+13]. It appears that the evaluation of quality considered by different viewers under different conditions could not be the same. According to the second approach, image quality is defined as the appreciation of an observer about a photo which is based on 2 notions : image aesthetic and image naturalness. More specifically, image aesthetic is the measure of how aesthetically a photo fulfills the observer's expectation (see Fig. 1.1(5) and Fig. 1.3) while the definition of image naturalness is both related to artifacts induced by some image processing algorithms and to the individual feeling about how a picture matches with image memory [LE+20] (see Fig. 1.1(6) and Fig. 1.5).

FIGURE 1.2 – Examples of intrinsic properties. The first photo has higher resolution than the others while the color depth of the third photo is shallower than the two first ones.

As a matter of fact, the main subject of the thesis is image quality. Focusing on the 2 factors : image aesthetic and image naturalness, we are not going to consider neither image content nor intrinsic image properties.

1.1 Image aesthetic and image naturalness

Image quality is the main subject of this thesis. There are 3 main aspects referring to image quality : intrinsic properties, image aesthetic and image naturalness (cf. Fig. 1.1). As presented before, intrinsic properties reflecting technical characteristics of images were the first historical concern about image quality but their influence on image quality has been insignificant because of the development of technology. In contrast, image quality depends on human expectations and viewers demand a photo that is not only faithful but also beautiful. Over time image aesthetic and image naturalness have become the 2 most important factor defining image quality. Although they have been identified as research topics for over 3 decades, they are still challenges because image aesthetic as well as image naturalness are related not only to universal validity referring to correctness or claims on viewers' understanding but also to subjectivity based on an individual feeling of pleasure or displeasure. The 2 contradictory principles of the 2 factors defining image quality refer to a more general debate "Is IQA subjective or universal ?".

1.1.1 Image aesthetic

Image aesthetic is an abstract notion dealing with how an observers feel about image appearance. This notion is related to what happens in viewers' mind when they view a photo (see examples in Fig. 1.3). The questions of how a photo is captured as well as how a viewer enjoys and criticizes the photo lead to the formation of photography. Photography is the art of creating durable images by capturing light. Although aesthetic opinions are not the same for each viewer, photography rules have been introduced based on describable aesthetic aspects in order to help taking a better photo. However, not all aesthetic aspects are describable so it does not mean that following photography rules always produces a high aesthetic photo and on the contrary a beautiful photo might not follow those rules. In photography, there are global rules referring to universal aesthetic aspects for all kind of images and particular rules formed for a particular image category. There are some aesthetic characteristics good for photos of a specific category and not good for other image categories. In particular, there are two image categories having opposite photography rules : large field images and close-up images. Looking at Fig. 1.4, the photos in the first row are large field images (images of a large field scene taken with a long distance from the camera to the scene) while the second row contains close-up images (images focusing on close-up objects captured with a short distance from the camera to the objects). On the side of large field images, photographers often follow the landscape rule to take large field photos. According to landscape rule, a photo is split in 3 regions (see examples in Fig. 1.4) and the compositions in a large field photo are usually arranged in those regions. In the first and the third photos, the horizon line is used to separate

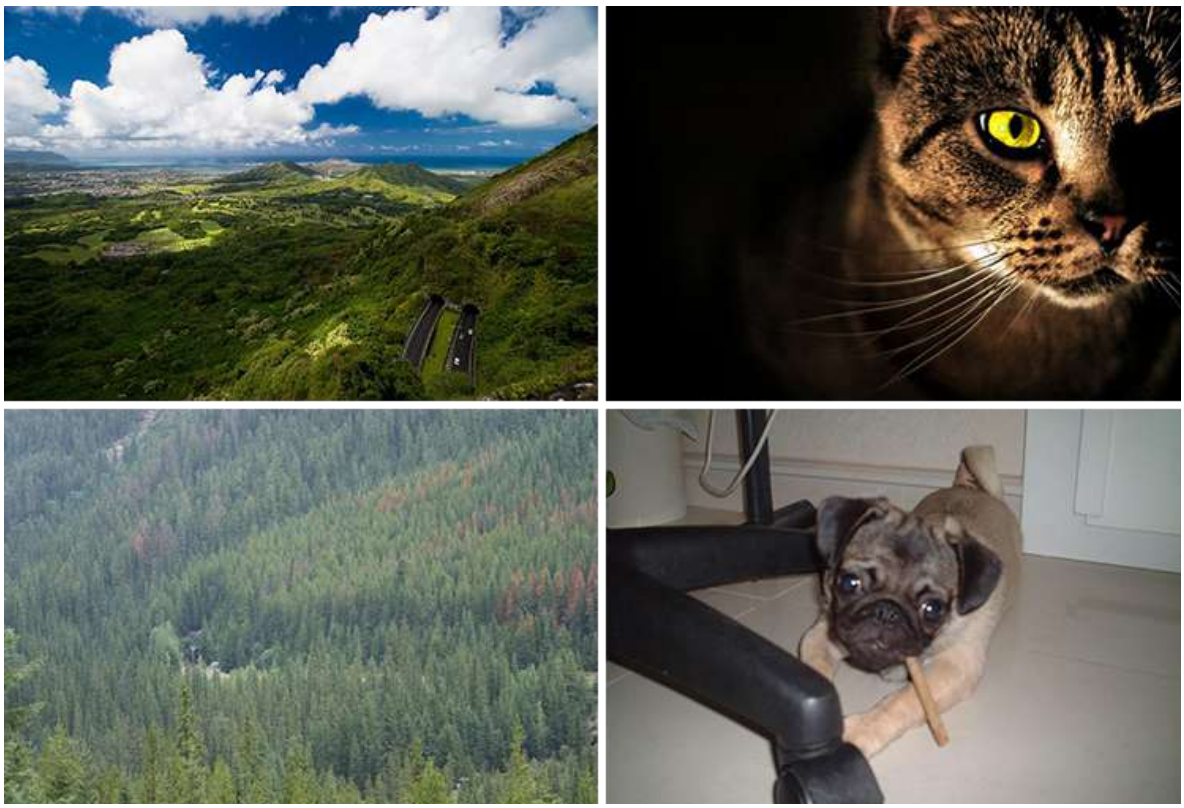


FIGURE 1.3 – Image aesthetic illustrations. Examples of high aesthetic images are in the first row while the second row contains examples of low aesthetic images.

the land and sky regions while in the second photo, the floor of the Colosseum is matched with the line of the landscape rule. In general, photographers usually exploit lines in scenes to separate regions. Another technique applied for large field images is to exploit patterns. For example, the patterns of grass, entrances and buildings are used in Fig. 1.4(a), (b) and (c) respectively. With regard to camera settings, large field images are taken with a long distance from the camera to the scene so photographers often set a small aperture setting to get a deep depth of field while a short focal length is chosen to take a wide scene.

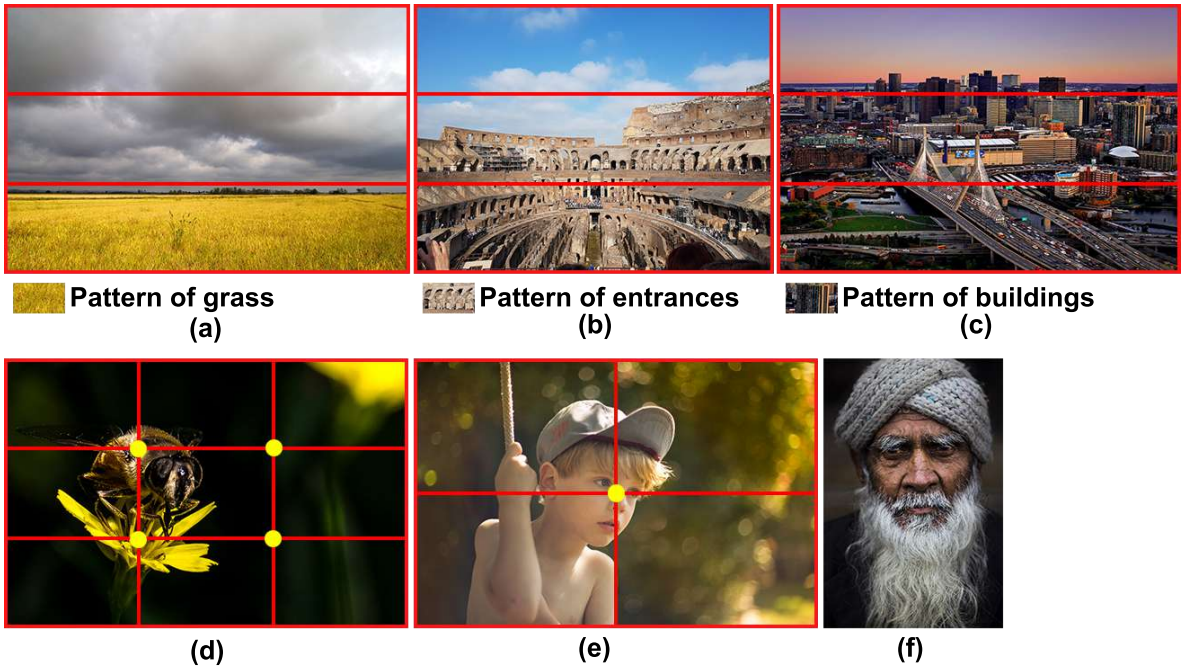


FIGURE 1.4 – Examples of large field and close-up images.

On the other side, the rule of thirds and the "center focus" rule are used when taking close-up images (see Fig. 1.4(d) and (e) respectively). While the rules for large field images focus on regions and lines, the rules for close-up images focus on points. According to the rule of thirds and the "center focus" rule, observers pay more attention to some special points (yellow points in Fig. 1.4(d) and (e)) so the main content of photos should be placed at (or near) those points. Photographers often use the effect "low depth of field" to blur the background in order to highlight the main content in a close-up image so the role of the background in close-up image aesthetic is insignificant. Thus, another rule applied for close-up photos is "fill the frame". Looking at Fig. 1.4(f), almost the whole photo frame is covered with the portrait of the old man and the background areas in this photo are smaller than the background in the 2 first close-up photos. However, this rule is rarely applied when taking large field images. It is opposite to the camera settings for taking large field images, a high aperture setting and a long focal length are often set in order to focus on close-up objects and to get a low depth of field when taking a close-up photo.

Image Aesthetic Assessment (IAA) is based on aesthetic features and those features are

somehow similar to photography rules so there are some universal features able to applied for all images but there are also particular features for a specific image category that could not be applied for the other image categories. Therefore there is an implicite assumption that evaluating image aesthetic under different views, criteria for each image category and image classification can improve the IAA performance.

Besides, some regions in an image are sharper than the others and some regions are salient because of their contrasted colors. Those regions seem to get more viewers' attention than the others. Looking at Fig. 1.4, in one hand, the close-up images have sharp object regions (the bee and the flower in the first photo, the boy and the old man in the second and the third ones respectively) and it appears that the color contrast between the background and the foreground is significant. On the other hand, the grass field, Colosseum and city scenes are the main content of the 3 large field images and viewers might be attracted more by those regions. Although the sky regions have less details and they are not the main content in those photos, their influence on aesthetic quality of the large field images is significant so the role of those regions is not like the role of the background in close-up images. There is an assumption that the aesthetic quality of an image is more related to the aesthetic quality of those regions than on the aesthetic quality of the whole image. This relation might be not the same for all images and it depends on the category of the considered photo.

To sum up, image aesthetic is one of the 2 main factors defining image quality and it is currently the most important concern about image quality so image aesthetic is going to be studied in this thesis. Besides, it appears that the ways of taking, composing or looking at an aesthetic picture of a large field scene and a close-up field scene are very different. This is why in our work, we are going to consider both categories of images for the purpose of IAA. Additionally, sharp and / or high contrasted color regions attract more viewers' attention than other regions so image aesthetic might have a stronger relation with salient regions than the whole photo frame. Thus, the role of region segmentation is going to be investigated for IAA.

1.1.2 Image naturalness

Image naturalness is a difficult notion to define. It is both related to artifacts and to the matching between the image appearance and the knowledge of reality stored in viewers' memory (see examples in Fig. 1.5). On the one side, image naturalness is affected by strong unnatural clues detected by viewers' eyes so the unnaturalness feeling comes from annoying artifacts induced by camera sensors, image processing algorithms (compressing, tone-mapping,...), image format, file transfer,... Those signs could be from imperceptible to obvious and a photo with annoying artifacts is considered as an unnatural photo (see bottom right photo in Fig. 1.5).

On the other side, the feeling of naturalness and unnaturalness comes from viewer's experience and memory (see bottom left photo in Fig. 1.5). When viewing a photo, observers compare the scene in the photo to reality retrieved from their memory (what they have seen) to find differences and similarities, so the feelings depend on individual factors. For example,

FIGURE 1.5 – Naturalness / unnaturalness illustrations. Examples of natural images are in the first row while the second row contains examples of unnatural images.

looking at Fig. 1.6, it appears that the colors of leaves in autumn are various and the colors change under different brightness conditions. A color might be familiar to some viewers but it could be not faithful to the others. In this case, some viewers consider colors that they have never seen as unnatural because they already fixed a color range for autumn leaves in their mind.

FIGURE 1.6 – Examples of various colors of leaves in autumn.

Although image aesthetic is the most important concern about image quality, it is not enough to precisely define image quality. A high aesthetic photo might not be a high quality photo if it is not faithful to viewers. Image naturalness is the counterpart of image aesthetic to define precisely image quality. Therefore image naturalness is studied in the thesis.

1.2 Objectives and outline of the thesis

Objectives : in this thesis, we are going to focus on image aesthetic and image naturalness which are two factors involved in image quality. Image intrinsic properties are not going to be considered (cf. Fig. 1.1).

More precisely, the first aspect (image aesthetic) is going to be investigated for two specific categories of images which are large field images and close up field images. Based on the assumptions that image categories and consequently region segmentation (salient areas and background) might have an influence on image aesthetic, our first contribution is to study image aesthetic assessed especially on images with and without pre-processing (image classification and region segmentation). Although there are many researches about Image Aesthetic Assessment (IAA), they mainly focus on developing methods for image aesthetic judgment while the question "how is the influence of pre-processing operations in IAA ?" has not been answered. In this study, the role of image classification and region segmentation in IAA is investigated first. Based on the estimated role, our second contribution is to present a new IAA model involving image classification and region segmentation (see Fig. 1.7).

The second objective is to study image naturalness of tone-mapped images. There are few researches about this topic so there are many questions that need to be answered. Our contributions to that research domain are twofold : proposing subjective experiments and developing objective metrics for Image Naturalness Assessment (INA) (see Fig. 1.8).

Last but not least, the relations and the correlations between image aesthetic and image

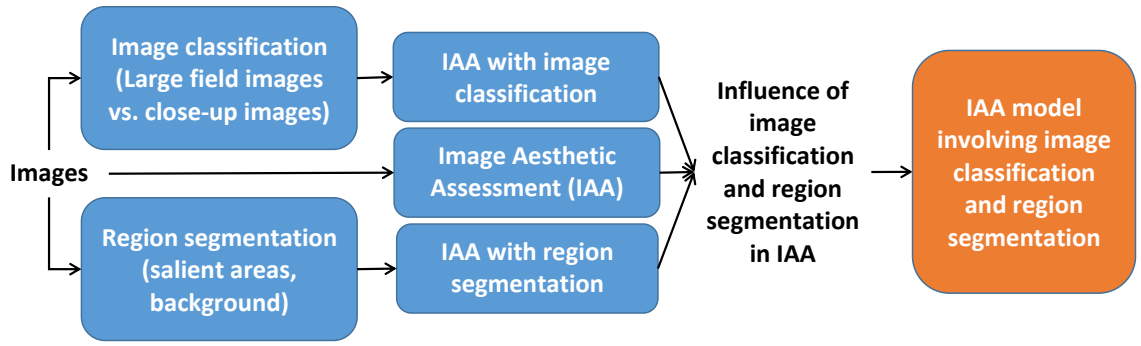


FIGURE 1.7 – Overview of the image aesthetic study.

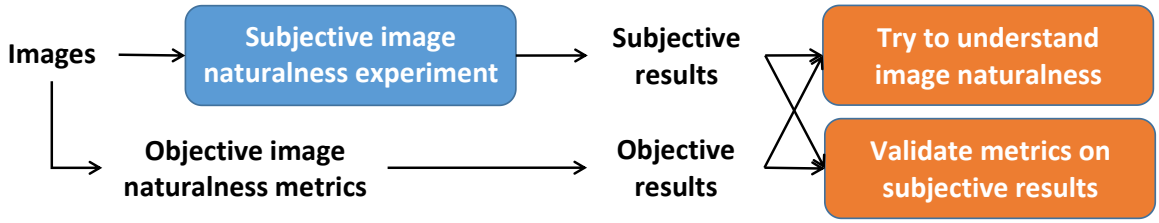


FIGURE 1.8 – Overview of the image naturalness study.

naturalness are investigated.

Outline : The main content of the thesis is organized in 4 chapters.

Chapter 1 describes the two proposed pre-processing methods for IAA including Region Of Interest Extraction (ROIE) methods and Large field / Close-up Image Classification (LCIC) methods. Region Of Interest (ROI) in this thesis is defined based on sharpness and color information. A handcrafted feature based ROIE algorithm and deep learning based ROIE methods are evaluated and compared. In a second part, various types of features including Exchangeable Image File Format (EXIF) features, handcrafted features and learned features are compared to address the LCIC task and provide an efficient classification algorithm.

Chapter 2 presents the study of IAA based on image classification and region segmentation. Firsly, the performances of IAA with and without LCIC and ROIE are considered to evaluate the role of image classification and region segmentation in IAA. Secondly, according to the estimated influence of image classification and region segmentation, an IAA model based on LCIC and ROIE is proposed. Both handcrafted features and learned features are exploited and validated on subjective data to evaluate the efficiency of the 2 types of features in the IAA task.

Chapter 3 describes a subjective experiment organized in laboratory environment to collect human evaluations about image naturalness of tone-mapped images. Based on the subjective data obtained from the experiment, different objective metrics are validated for

the INA task. Beside that, the definition of image naturalness is clarified and factors affecting image naturalness are explained and discussed. Additionally, the relations and the correlations between image aesthetic and image naturalness are addressed.

Chapter 4 summarizes the contributions of the thesis. Beside that, perspectives about future works are drawn up.

Chapitre 2

Pre-processing for image aesthetic assessment

Sommaire

When a viewer looks at a photo, some regions receive more attention from the viewer than other regions. Those regions are defined as Regions Of Interest (ROI). As a matter of fact, there is an implicite assumption that the aesthetic quality of an image is more related to the aesthetic quality of the ROI in this image than on the aesthetic quality of the whole image. Looking at Fig. 2.1, the ROIs (represented by white regions in the second row) of the photos in the first row are more salient and attract more viewers' attention than the background (represented by the black regions in the second row). The first contribution of this chapter is to propose 2 ROI Extraction (ROIE) methods based on sharpness and color information : the first method being handcrafted based and the second one being deep learning based. Secondly, following the idea that photos in different categories (human, flower, animal, landscape,...) are taken with different photographic techniques, image aesthetic should be evaluated in a different way for each image category. Large field images and close-up images are 2 typical categories of images with opposite photographic rules so we want to investigate the intuition that prior Large field / Close-up Image Classification (LCIC) might improve the performance of IAA. To do that, there is a need of LCIC algorithm development, so the second contribution

of this chapter is to propose and compare LCIC algorithms based on different types of features for the classification purpose : Exchangeable Image File Format (EXIF) features, handcrafted features and learned features.

FIGURE 2.1 – Examples of close-up images (on the left), large field images (on the right) and the corresponding ROIs (second row).

2.1 Introduction

When looking at an image, sharp regions and salient color regions often attract more viewers' eyes while background areas often get less viewers' attention. Thus, sharpness and color saliency are 2 factors defining the Region of Interest (ROI) we are looking for. In the first row of Fig. 2.1, the left photo is a close-up image of tulip flowers while on the right side, the photo is the large field scene of a tulip field. Although viewers pay more attention to some ROIs in both cases, the roles of the background areas in the aesthetic quality of the 2 photos are different. In the close-up photo, the blur background and the high contrasted colors between the ROIs and the background are exploited to highlight the sharp and high contrasted color flowers. On the contrary, although the main objects in the right photo are the colorful tulip field and the windmills, the roles of the blue sky and white clouds are significant in the aesthetic quality of the image because the whole image is usually considered when assessing aesthetic of large field images. Thus, the role of the background in the large field photo is different from that of the close-up photo. According to the above analysis, we can see 2 points. Firstly, ROIs have a significant influence on image aesthetic but this influence is not the same for each image category. Secondly, different criteria should be considered when assessing aesthetic quality of different image categories. In other words, performing image classification and ROI Extraction

(ROIE) before Image Aesthetic Assessment (IAA) can enhance the IAA performance. The two image categories focused here are large field images (images of a large field scene taken with a long distance from the camera to the scene) and close-up images (images focusing on close-up objects captured with a short distance from the camera to the objects) because of the obvious differences of photographic rules and aesthetic evaluation criteria between them. Moreover those both categories contain a huge amount of possible images.

In this chapter, there are two main contributions regarding pre-processing for IAA. The first one is to study ROIE. An ROIE algorithm using the combination of sharpness and color contrast information and a deep model extracting ROIs are introduced. The second contribution is to consider different types of features including Exchangeable Image File Format (EXIF) features, handcrafted features and learned features to perform the Large field / Close-up Image Classification (LCIC) task. The performances of LCIC based on each feature set are compared in terms of accuracy and computational complexity [LE+19].

2.2 Region Of Interest Extraction (ROIE)

2.2.1 State of the art

There are many ways to extract ROIs. The first way is to consider image sharpness because viewers are often attracted by sharp and clear regions when viewing a photo. Following this idea, from an input image, Luo et al. [LT08] use blurring kernels, horizontal and vertical derivatives to compute sharpness information. Each pixel is labelled as blur or clear and the ROIs are considered as the rectangular regions with the highest sharpness values. Their results are not really accurate because the shape of any ROI is not always rectangular. Reusing Luo's sharpness calculation, Tang et al. [TLW13] propose first to segment the input image into super-pixels (groups of neighboring pixels having similar colors) [Ach+12] and then the labels of neighboring pixels are used to improve the precision of ROIE. A super-pixel is determined as belonging to an ROI if over half of its pixels are labelled as clear. Tang's ROIE is better than Luo's one since the extracted ROIs' shapes look more visual (more similar to the shape of objects in photos instead of rectangular regions). In [ASG15], Aydin et al. use an edge stopping pyramid to blur the input image multiple times. By considering the differences between the blurred versions of the sequential pyramid levels, a sharpness map is computed first and the in-focus regions are then extracted based on it.

The second approach is based on the fact that regions with salient and high contrasted colors often get more viewers' attention. Jiang et al. [Jia+11] introduce a segmentation method integrating color saliency and object-level shape prior to extract ROIs. The method is based on 3 main characteristics of salient objects : differences from the surrounding context, a center location and a well-defined closed boundary. In [Per+12], Perazzi et al. use color contrast and color distribution to estimate the color saliency level of each super-pixel. Color variations, spatial frequencies, structure and distribution of image segments are considered in their study. In [Che+13], a color saliency prediction algorithm exploiting global color uniqueness and

color spatial distribution is proposed. Fu et al. [Fu+13b] introduce an ROIE method using a combination of global color contrast and Harris convex hull. A global propagation procedure via geodesic distance is the primary key of their method. In [ZZC13] an algorithm using the combination of color dissimilarity with background prior for color saliency level computation is proposed. That combination achieves a higher accuracy than some previous methods if salient regions are located at the center of images. In [Ton+15], exploiting both weak and strong models, a salient object detection method combining color saliency and bootstrap learning to extract salient regions is proposed. A weak saliency map is constructed first based on image priors to generate training samples for a strong model. Then, the strong classifier is learned to detect salient pixels from images directly. In [Che+15], a color saliency detection method analyzing color histogram and spatial information-enhanced region based contrast is proposed.

Beside handcrafted methods, deep learning based methods have been also developed for region detection and saliency prediction [LY15] ; [Zha+15] ; [Cor+16] ; [LH16]. In [Zha+15], Convolutional Neural Networks (CNNs) are used to modelize saliency of objects in images by considering both global and local contexts. Saliency features are extracted from 2 models, one trained on the global context and the other trained on the local contexts. Both feature types are then used for color saliency computation. Li et al. [LY15] propose to use CNNs to learn saliency features from multiscale images for visual recognition tasks. Different visual saliency maps are generated from multiscale images coming from an original one. Those maps are then combined to create the final saliency map. In [LH16], an end-to-end deep hierarchical network based on CNN for salient object detection is proposed. The first network learns global contrast, objectness, compactness features. Then a hierarchical recurrent CNN is used to hierarchically refine the details of saliency maps by integrating local context information. Cornia et al. [Cor+16] propose to predict viewers' attention on image pixel by using an CNN containing 3 main blocks : a feature extraction CNN, a feature encoding network and a prior learning network. That model extracts deep features from different levels of the CNN and combines them to predict eye fixations over the input image.

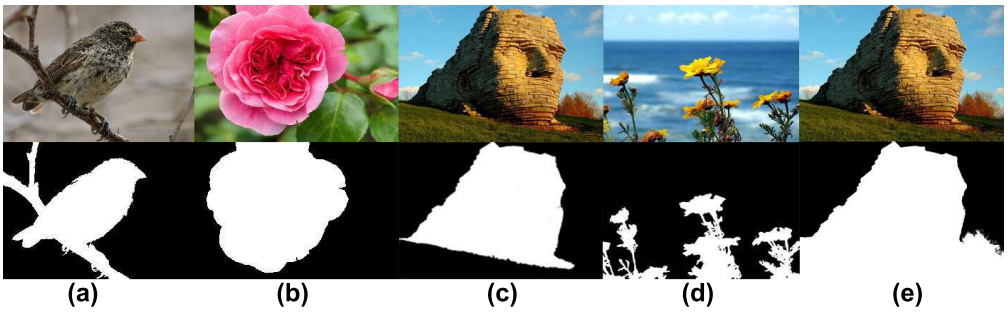


FIGURE 2.2 – Examples of different definitions of ROIs. The first row contains color images and the second row contains the corresponding ROI maps (a) ROIs defined by sharpness. (b) ROIs defined by color saliency. (c) ROIs defined as object regions (d)(e) Our ROI definition based on both sharpness and color saliency.

In our work, ROIs are defined as regions attracting viewers' attention by both sharpness

and color saliency factors (see Fig. 2.2(d) and Fig. 2.2(e)). They are not only sharp regions or only regions with high color saliency levels or regions containing objects (see Fig. 2.2(a), Fig. 2.2(b), Fig. 2.2(c) respectively).

2.2.2 Handcrafted ROIE method

The main idea of this method is based on the fact that observers pay more attention on sharp regions or regions having salient colors. Therefore, the first step is to estimate the sharpness of all regions in the image. In the second step, the color saliency levels are computed. The estimated sharpness and color saliency levels are combined to form the ROI map in the last step.

2.2.2.1 Sharpness map

In-focus regions are defined as the regions focused by photographers when taking a photo. Normally, the in-focus regions are sharper than the other regions so sharpness information is the primary key to detect them. Sharpness is a combination of resolution and acutance. It quantifies the variations in gray scale values between neighboring pixels. Professional photographers often use sharpness as one of the main factors to distinguish between ROIs and background. In [Cre+07] ; [ASG15], they point out that when blurring a photo, the neighboring pixels' values converge to the same gray level. The gray levels of pixels in a sharp image change significantly when the image is blurred. This change is much weaker when a blurred image is re-blurred (see Fig. 2.3 for examples). To extract in-focus regions, a sharpness estimation method based on the combination of Aydin's clearness map [ASG15] and multi-scale super-pixels is introduced.

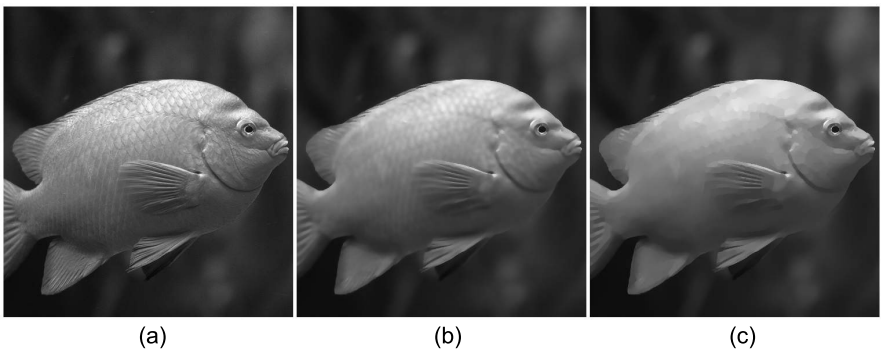


FIGURE 2.3 – The changes of gray level pixels after blurring and re-blurring. (a) original image, (b) blurred image, (c) re-blurred image.

The first step of the sharpness estimation process is to calculate Aydin's clearness map. The key idea of the clearness map computation is to consider the variations of differences in

gray scale after blurring the image. A k-level edge-stopping pyramid [ASG15] is built by using the bilateral filter [TM98]. The first pyramid level L_0 is the image in gray scale while the higher levels are defined as :

$$L_i = f_b(L_{i-1}, s_i) \tag{2.1}$$

where f_b is the bilateral filter. In this work, k is set to 10. The kernel size at the i^{th} level is $s_i \times s_i$ where $s_i = round(3 \times 1.1^i) \times 2 + 1$. The clearness map is then calculated as the sum of absolute differences between subsequent pyramid levels as :

$$M^{cl} = \sum_{i=1}^{k} |L_i - L_{i-1}| \tag{2.2}$$

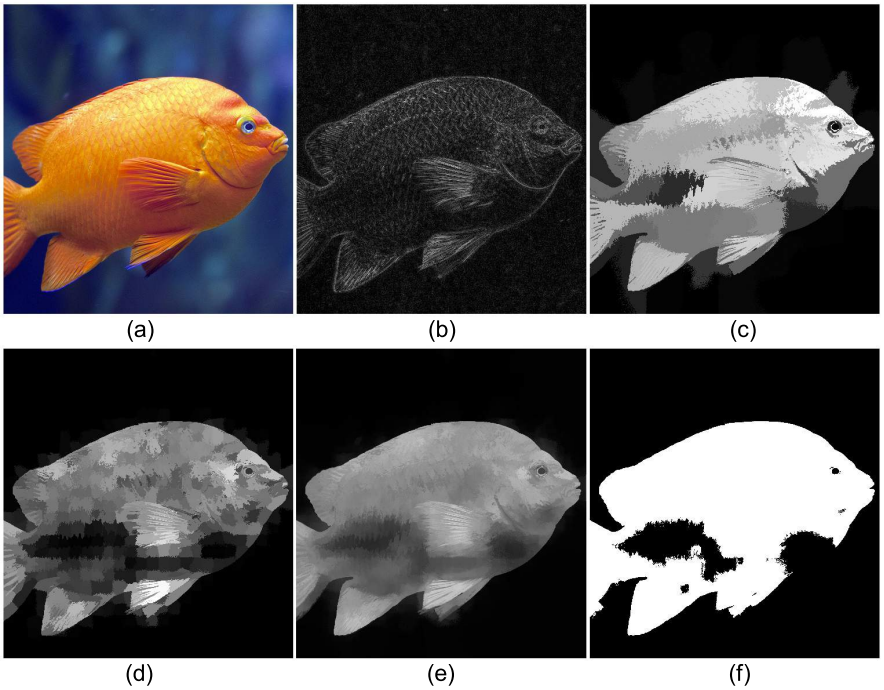


FIGURE 2.4 – Sharpness map computation process. (a) original image, (b) Aydin's clearness map, (c) sharpness distribution at level 2, (d) sharpness distribution at level 5, (e) sharpness map, (f) in-focus map.

Aydin's clearness map only gives a rough estimation of the sharpness. As an example, the Aydin's clearness map is presented in Fig. 2.4(b), the detected sharp pixels are located mainly on edges while viewers often pay attention to the whole regions containing sharp details instead of all small sharp details. In this work, our sharpness map is calculated by spreading the clearness values over super-pixels. In the next step, n multi-scale super-pixel levels are determined. At the i^{th} level, the color image is segmented into $i^2 \times \alpha$ super-pixels ($\alpha = 25$, $n = 10$ in this work). The sum of clearness values $s^{cl}_{i,j}$ of super-pixel P_j at the i^{th} level is

calculated as :

$$s_{i,j}^{cl} = \sum_{(x,y)\in P_j} M^{cl}(x,y) \tag{2.3}$$

After normalizing the $s_{i,j}^{cl}$ values to the range [0, 255], sharpness values of all pixels in each super-pixel P_j are set to $s_{i,j}^{cl}$ and the sharpness distribution map M_i^{sh} at the i^{th} level is obtained (see Fig. 2.4(c) and Fig. 2.4(d) for illustrations). The global sharpness map is then computed as :

$$M^{sh} = \frac{1}{n}\sum_{i=1}^{n} M_i^{sh} \tag{2.4}$$

Comparing Fig. 2.4(b) and Fig. 2.4(e), it appears that pixel values in our sharpness map are more consistent with the visual content of the image and with what people see as sharp regions. The contribution here is to exploit color and spatial information of pixels (super-pixels) to increase the precision of the sharpness estimation. The sharpness map is then binarized by applying Otsu's threshold [Ots79] to extract the in-focus regions. The in-focus map is the binarized version of the sharpness map (see Fig. 2.4(f)).

2.2.2.2 Color saliency map

Beside the sharpness factor, color contrast is another important factor attracting viewers' attention. Previous researches mainly focus on color differences between all the regions in the image [Per+12] ; [ZZC13] ; [Fu+13a] ; [PSh15] ; [Liu+17]. Based on Liu's idea [Liu+17] and Zheng's idea [ZZC13] about using background and foreground priors and Perazzi's idea [Per+12] about using color uniqueness, we propose a color saliency estimation algorithm combining both the background, foreground priors and the color uniqueness. Salient regions in this work are defined as regions having colors similar to the colors of the in-focus or central regions (central regions attracting more attention than out of central regions - regions near photo edges) and different from the colors of the other regions (out of focus regions and out of central regions). After identifying the in-focus regions as in the previous part, a mask is initialized based on the in-focus regions and the center region :

$$M^{msk} = M^{inf} \cup M^{cen} \tag{2.5}$$

where M^{inf} is the in-focus map. M^{cen} is a binary image in which there are a white center rectangular region of size $0.6w \times 0.6h$ and the other black background regions (w and h are the width and the height of the image). The color saliency M_i^{cs} of super-pixel P_i is estimated by using color differences between P_i and all out-of-mask super-pixels and color similarities between that super-pixel and all in-mask super-pixels as :

$$M_i^{cs} = \frac{\sum_{P_j\in R_{oom}} d_{i,j}^{rgb} \times w_{i,j}^{p}}{\parallel R_{oom} \parallel} - \frac{\sum_{P_j\in R_{inm}} d_{i,j}^{rgb} \times w_{i,j}^{p}}{\parallel R_{inm} \parallel} \tag{2.6}$$

where R_{oom}, R_{inm}, $\parallel R_{oom} \parallel$, $\parallel R_{inm} \parallel$ are the out-of-mask, in-mask regions and the number of super-pixels in those regions respectively. $d_{i,j}^{rgb}$ is the color distance between the center pixels of

super-pixels P_i and P_j in RGB color space. Gaussian weight $w^p_{i,j}$ is calculated via super-pixel center positions by applying the following formulas :

$$d^{rgb}_{i,j} = \sqrt{(r_i - r_j)^2 + (g_i - g_j)^2 + (b_i - b_j)^2} \tag{2.7}$$

$$w^p_{i,j} = \frac{1}{z^p_i} e^{-\frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{2\sigma_p}} \tag{2.8}$$

where x_i, y_i, r_i, g_i, b_i are the coordinates and red, green, blue intensities of the center pixel in P_i. σ_p is the number of super-pixels in the image. The normalization factor z^p_i ensures $\sum_{P_j \in R_{oof}} w^p_{i,j} = 1$.

Pixel values in M^{cs} are normalized to the range $[0, 255]$ and the Otsu's threshold is applied on M^{cs} to create an update of the mask M^{msk} and a new cycle starts. After performing this process 3 times, the final color saliency map M^{cs} is obtained (see examples of color saliency maps in Fig. 2.5(c)).

2.2.2.3 Region of interest map

Looking at Fig. 2.5, it appears that sharpness is the main factor attracting viewers' attention in the two first rows. In contrast, the dominant criterion emphasizing the ROIs is the color saliency in the two last rows. For the three middle rows, both sharpness and color saliency have significant roles in highlighting the ROIs. Obviously, the role of sharpness and color saliency factors in defining ROIs is not the same for all images. Thus, if only one of them is considered, it will not be sufficient to extract right ROIs. An algorithm combining sharpness and color saliency factors to extract ROIs is presented in this part.

The spatial distribution of pixel values is the key for estimating the roles of sharpness and color saliency factors in attracting viewers' attention. We introduce a method presenting the distribution of pixel values by using a rectangle. Given a gray image (either a sharpness map or a color saliency map) I, the coordinates of the center point of the rectangle are first determined as :

$$x_c = \frac{\sum_{x=1}^{w} \sum_{y=1}^{h} I(x,y) \times x}{\sum_{x=1}^{w} \sum_{y=1}^{h} I(x,y)} \tag{2.9}$$

$$y_c = \frac{\sum_{x=1}^{w} \sum_{y=1}^{h} I(x,y) \times y}{\sum_{x=1}^{w} \sum_{y=1}^{h} I(x,y)} \tag{2.10}$$

These coordinates are then used to calculate the deviations as :

$$d_l = \frac{\sum_{x=1}^{x_c} \sum_{y=1}^{h} I(x,y) \times |x - x_c|}{\sum_{x=1}^{x_c} \sum_{y=1}^{h} I(x,y)} \tag{2.11}$$

$$d_r = \frac{\sum_{x=x_c}^{w} \sum_{y=1}^{h} I(x,y) \times |x - x_c|}{\sum_{x=x_c}^{w} \sum_{y=1}^{h} I(x,y)} \tag{2.12}$$

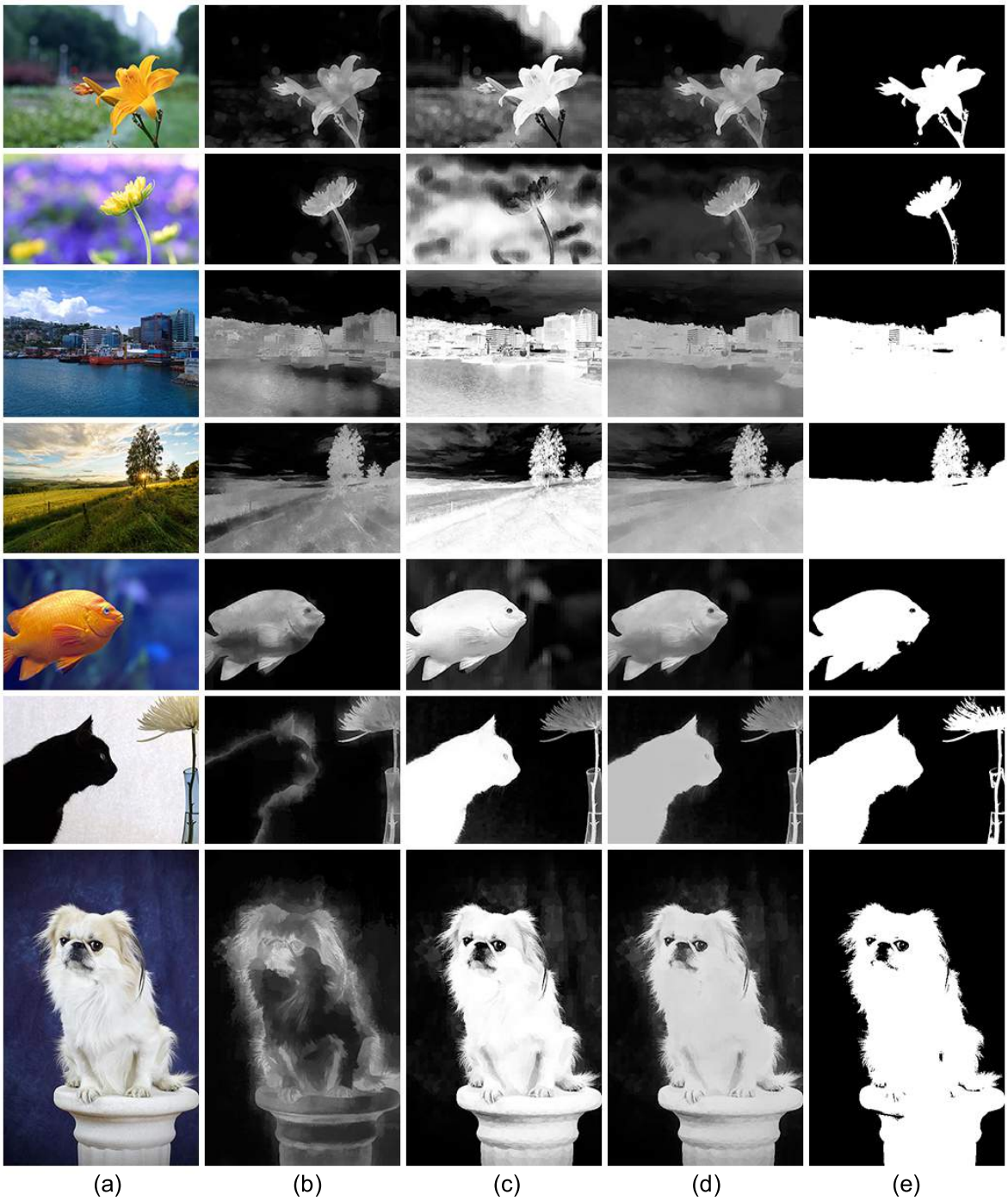


FIGURE 2.5 – ROI map computation process. (a) original images, (b) sharpness maps, (c) color saliency maps, (d) ROI maps. (e) binarized ROI maps.

$$d_t = \frac{\sum_{x=1}^{w}\sum_{y=1}^{y_c} I(x,y) \times |y - y_c|}{\sum_{x=1}^{w}\sum_{y=1}^{y_c} I(x,y)} \tag{2.13}$$

$$d_b = \frac{\sum_{x=1}^{w}\sum_{y=y_c}^{h} I(x,y) \times |y - y_c|}{\sum_{x=1}^{w}\sum_{y=y_c}^{h} I(x,y)} \tag{2.14}$$

where d_t, d_r, d_b and d_l are the top, right, bottom and left deviations respectively. The rectangle R_I representing the distribution of pixel values in the image I is illustrated by the red rectangles in Fig. 2.6(b) and Fig. 2.6(c). The coordinates of the top left and bottom right points of R_I are computed as :

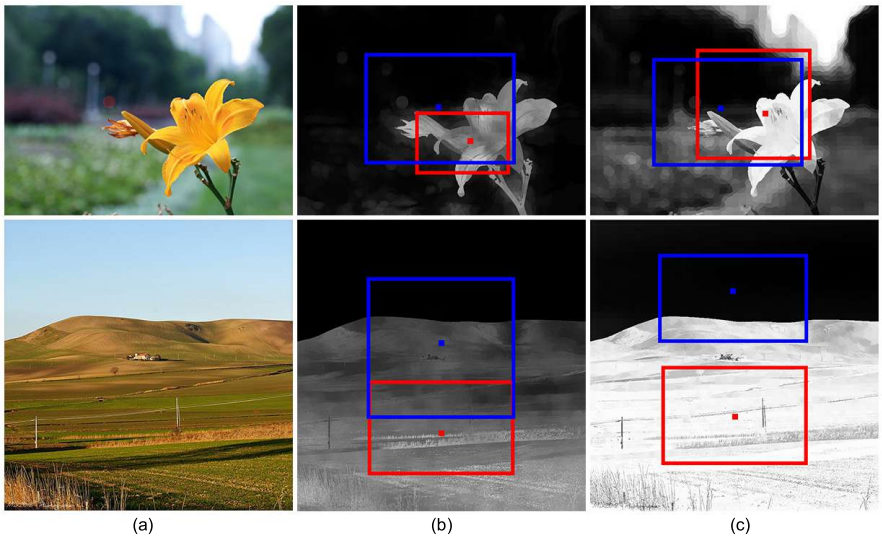


FIGURE 2.6 – Examples of rectangles representing the distribution of pixel values. (a) original images, (b) sharpness maps, (c) color saliency maps. Red rectangles represent the distributions of pixel values in those images while blue rectangles reflect the distributions for the corresponding video inverted images.

$$x_{tl} = x_c - d_l \tag{2.15}$$

$$y_{tl} = y_c - d_t \tag{2.16}$$

$$x_{br} = x_c + d_r \tag{2.17}$$

$$y_{br} = y_c + d_b \tag{2.18}$$

The distribution rectangle concept is then used to estimate the influences of sharpness and color saliency factors in attracting viewers' eyes. Comparing the red and blue rectangles in the first row of Fig. 2.6, there is a correlation between the size of the rectangle and the discrimination power of the data regarding viewers' attracting attention : The role is more significant when the size of the rectangle is smaller. If salient and un-salient regions are separated in opposite sides, the separation of them will be obvious as shown in the second row of Fig. 2.6. If the contrast is low, R_I and $R_{\neg I}$ will be large and intersect each other where

R_I is the rectangle representing the distribution of pixel values in the image I ($\neg I$ is the video inverted image of I). If the contrast is high, the size of R_I and $R_{\neg I}$ will be small and they could not intersect or the intersection could be insignificant because they are located in opposite sides (see the bottom images in Fig. 2.6(b) and Fig. 2.6(c)). Therefore, the sharpness and color saliency weights are computed as :

$$w_{sh} = \left(\frac{\parallel R_{\neg M^{sh}} \parallel}{\parallel R_{M^{sh}} \parallel + \parallel R_{M^{sh}} \cap R_{\neg M^{sh}} \parallel}\right)^2 \tag{2.19}$$

$$w_{cs} = \left(\frac{\parallel R_{\neg M^{cs}} \parallel}{\parallel R_{M^{cs}} \parallel + \parallel R_{M^{cs}} \cap R_{\neg M^{cs}} \parallel}\right)^2 \tag{2.20}$$

The values of w_{sh} and w_{cs} reflect the influences of sharpness and color saliency in highlighting ROIs. The proposed ROI map is calculated as :

$$M^{roi} = \frac{w_{sh} \times M^{sh} + w_{cs} \times M^{cs}}{w_{sh} + w_{cs}} \tag{2.21}$$

The binarized version M_b^{roi} of the ROI map M^{roi} is then obtained by applying the Otsu's threshold to extract the ROIs. In Fig. 2.5(c) and Fig. 2.5(d), examples of the proposed ROI map and the binarized ROI map are shown.

2.2.3 Deep learning based ROIE method

Beside handcrafted approaches, deep learning based approaches might be a promising solution for ROIE. In this part, 3 typical architectures are studied for ROIE. The 2 first models are designed based on a well-known architecture with 3 main components : encoding, transformation and decoding components while the third one is designed based on a traditional architecture with only convolutional blocks. The structures of the 3 models are presented in Fig. 2.7, Fig. 2.8 and Fig. 2.9.

In the 2 first models, the first component contains 3 blocks of convolutional layers (see Fig. 2.7(a)). In each block, a convolutional layer is connected to an instance normalization layer and it is activated by the ReLU function. The encoding component receives input color images of size 600×600 and passes the output to the transformation component. In the first model there are 5 residual blocks in the transformation component. The structure of a residual block is illustrated in Fig. 2.7(b) with 2 blocks of convolutional layers. The transformed data is then concatenated with the input data to create the output of the block. In the second model, the transformation block contains 10 convolutional blocks (see the structure of a convolutional block in Fig. 2.8(b)). The data transformed by the transformation component is passed through convolutional transpose layers of the decoding component and actived by a $Tanh$ activation function to generate the binary ROI maps. The difference between the 2 first models is in the transformation components : the first model uses residual blocks while the second one uses convolutional blocks. On the contrary, the third model includes convolutional blocks only. There are 8 convolutional blocks in the model and each block has a convolutional layer, an

instance normalization layer and an ReLU activation layer (see Fig. 2.9). The numbers of kernels in the blocks are 24, 48, 96, 192, 96, 48, 24 and 1 respectively. The input layer and the output layer of the third model are similar to those of the 2 first models. The points we want to clarify here are "Among these 3 typical architectures, which one is the best for ROIE ?".

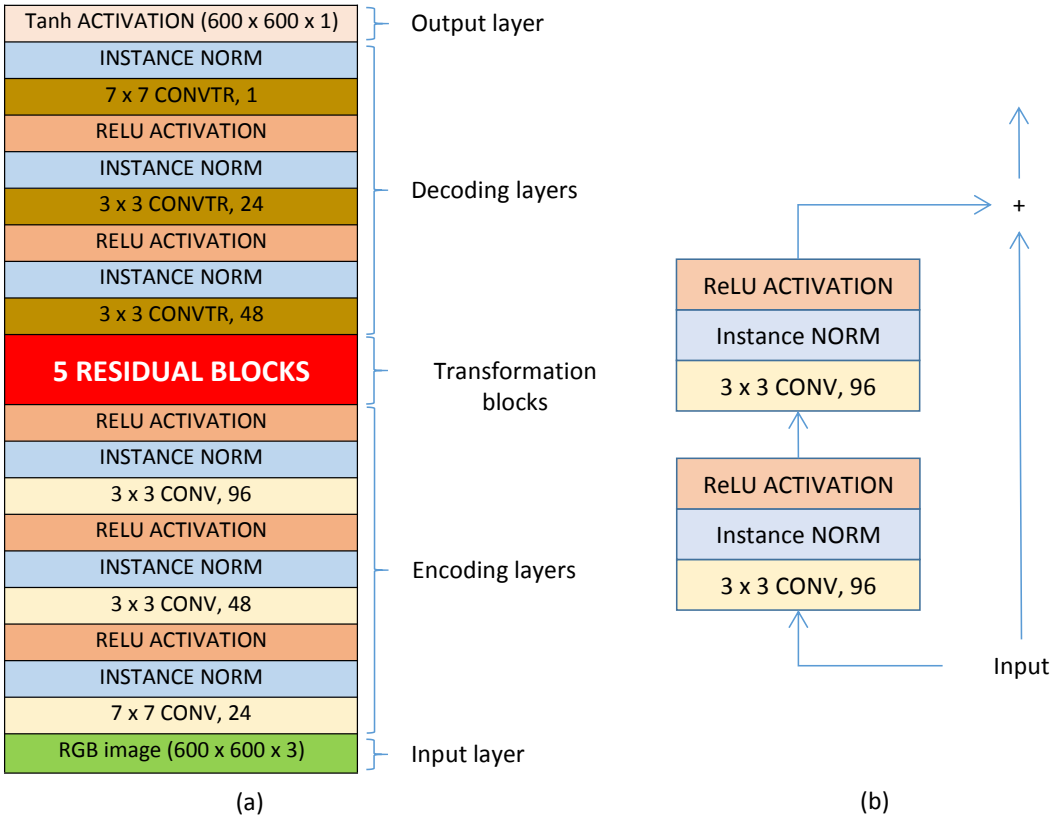


FIGURE 2.7 – The structure of the first model containing 3 main components : encoding component, transformation component (using residual blocks) and decoding component. (a) The structure of the model. (b) The structure of a residual block.

2.2.4 Experiment and results

2.2.4.1 Dataset and setup

1156 images (406 images from the CUHKPQ dataset [TLW13] and 750 images from the Flickr.com website) are selected for the experiment. Following the ROI definition presented in 2.2.1, each image is associated to a binary ground truth produced by the authors. The blur

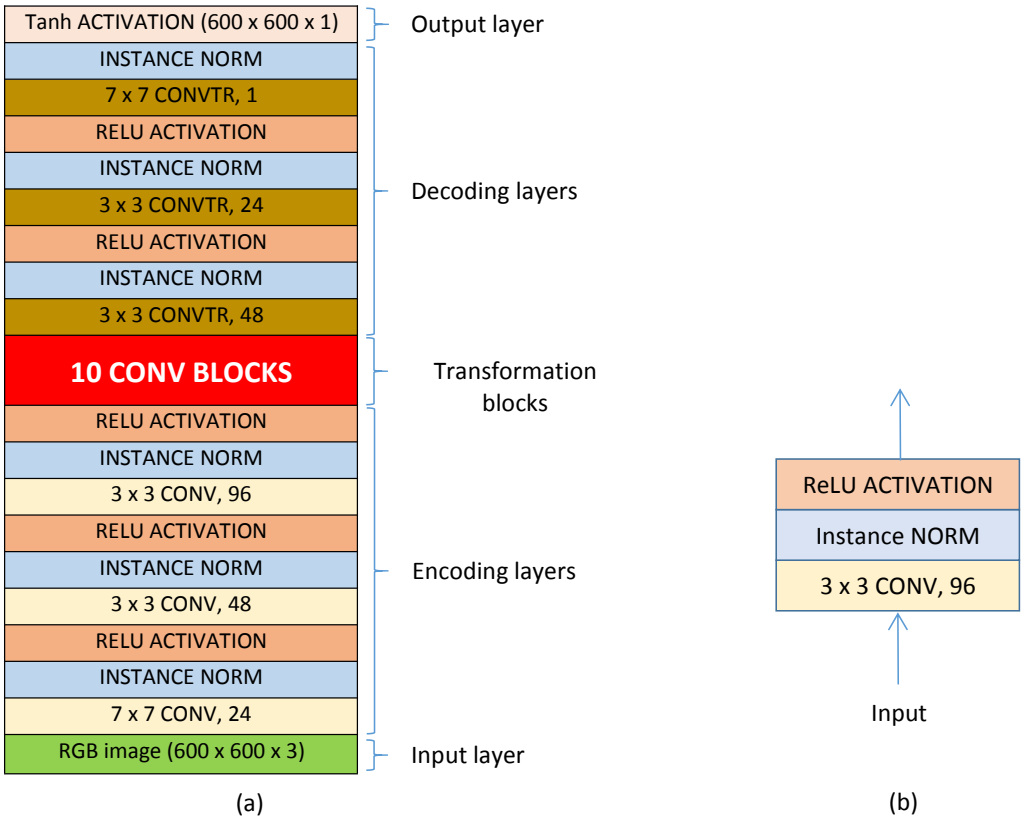


FIGURE 2.8 – The structure of the second model containing 3 main components : encoding component, transformation component (using convolutional blocks) and decoding component. (a) The structure of the model. (b) The structure of a convolutional block.

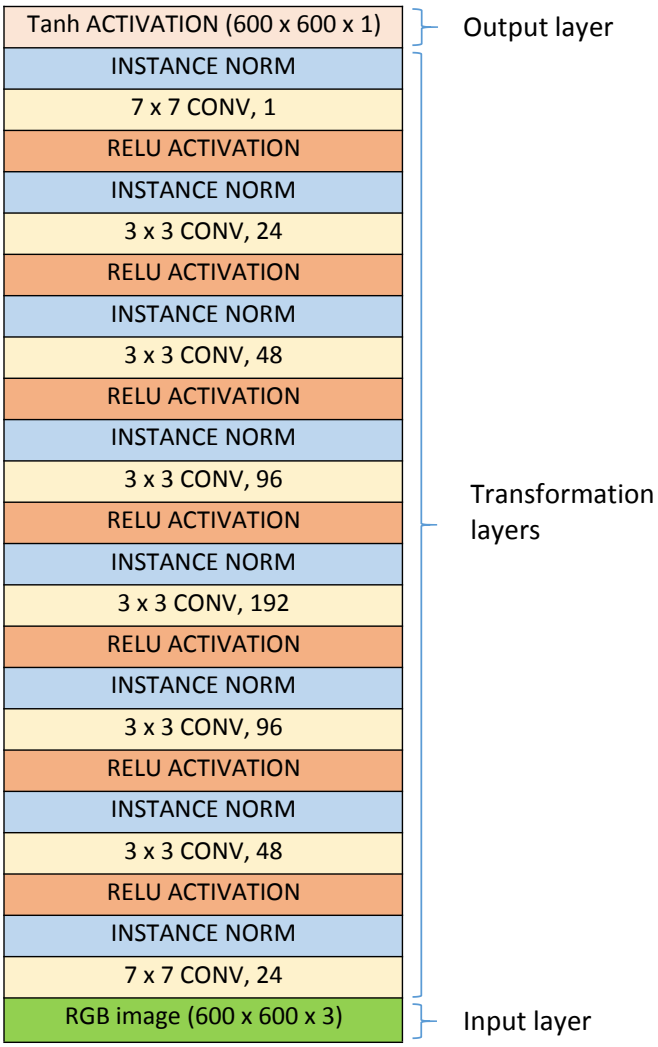


FIGURE 2.9 – The structure of the third model containing only convolutional blocks. There are 8 convolutional blocks with the numbers of kernels in the blocks are 24, 48, 96, 192, 96, 48, 24 and 1 respectively.

regions and unsalient color regions are considered as background (black regions in Fig. 2.10) while sharp, high contrasted color regions are determined as ROIs (white regions in Fig. 2.10). The proposed ROIE methods are evaluated on the dataset and they are compared with two methods based on sharpness information only (Aydin's [ASG15] and Tang's [TLW13] methods) and with two methods based on color contrast information only (Perazzi's [Per+12] and Zheng's [ZZC13] methods).

For the deep learning based approach, in order to train and test the deep models, the dataset is divided into 4 parts (each part contains 289 images). The models are trained 4 times. Each time, only one part is used for the test while the others are considered as the training set. To increase the number images in training sets (because training those deep models requires a big number of samples), a data augmentation process is applied. From an image, 200 augmented versions of size 600×600 are generated by flipping, re-scaling, padding, modifying brightness and shifting (see Fig. 2.10). In the training phase, the chosen optimizer is Adam optimizer and the loss function is the mean squared error function while the learning rate is set to 10^{-4}.

Five comparisons have been made to evaluate the methods. Firstly, based on the idea that sharp and clear regions attract more viewers' eyes, the proposed sharpness estimation method is evaluated and compared with 2 methods based on sharpness information (Aydin's and Tang's methods). Secondly, following the remark that high color saliency regions get more observers' attention than other regions, the comparison between the proposed color saliency map and 2 ROI maps based on color contrast information (Perazzi's and Zheng's color saliency maps) is performed. The third one is to compare the handcrafted ROI maps based on both sharpness and color information with the proposed sharpness maps and the proposed color saliency maps. The next comparison is for ROI maps generated by the deep models to find the best model for ROIE. The last comparison is between the handcrafted approach and the deep learning based approach for ROIE.

For a given map in gray scale, pixel values range from 0 to 255, except for Tang's ROI maps and ROI maps generated by the deep models (they are binary maps). The simplest way to compare those maps with the binary ground truth is to convert them to binary levels by applying a threshold. In this work, two thresholds have been considered. The first way is to use every threshold ranging from 0 to 255. The results are then used to form a precision recall curve. The Area Under Curve (AUC) is considered as the evaluation criterion. The second way is to choose a fixed threshold in which there are two options : Otsu's threshold selected based on the gray histogram and the adaptive threshold defined as twice the mean of pixel values [Ach+08]. After performing the experiments, we conclude that applying Otsu's threshold makes better results than applying the adaptive threshold so only results gained with Otsu's threshold are presented in this section. The evaluation criteria with a fixed threshold are precision, recall, F-measure and IoU that are defined in Table 2.1. The range of a metric X within the 95% confidence interval [Mit97] ; [DE96] is described as $X \pm I_X$. The interval I_X is calculated as :

$$I_X = z \times \sqrt{\frac{X \times (1 - X)}{N}} \tag{2.22}$$

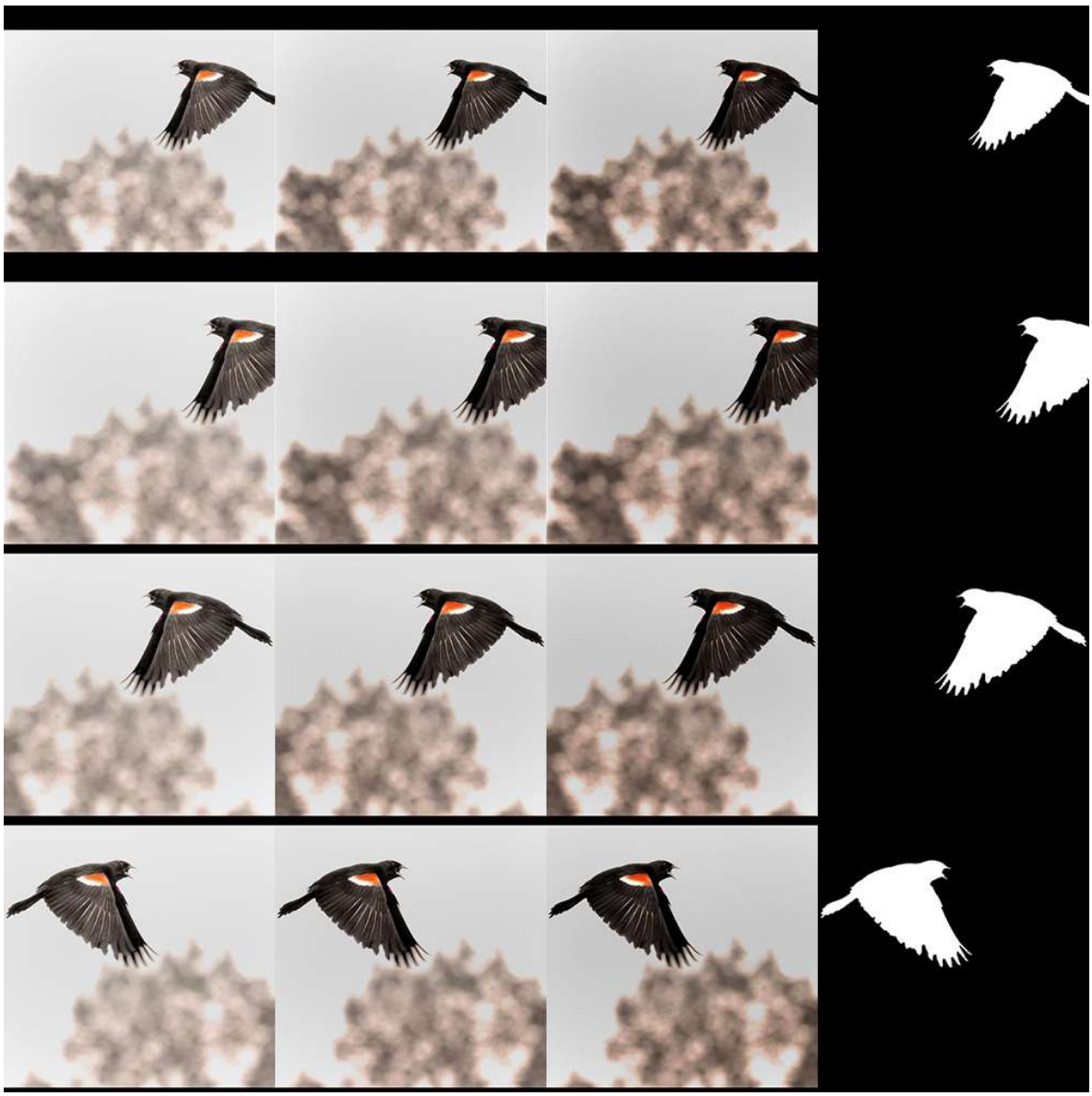


FIGURE 2.10 – Examples of data augmentation. The three left columns contain the augmented versions while the last column shows the ROI ground truth for the augmented versions in the corresponding row.

Evaluation criteria of ROI detection methods			
		Prediction	
		ROIs	Background
Ground truth	ROIs	tp	fn
	Background	fp	tn
Precision	$pr = \frac{tp}{tp+fp}$	$I_{pr} = z \times \sqrt{\frac{pr \times (1-pr)}{N}}$	
Recall	$re = \frac{tp}{tp+fn}$	$I_{re} = z \times \sqrt{\frac{re \times (1-re)}{N}}$	
F-measure	$F_\beta = \frac{(1+\beta^2) \times pr \times re}{\beta^2 \times pr + re}$	$I_{F_\beta} = z \times \sqrt{\frac{F_\beta \times (1-F_\beta)}{N}}$	
Intersection over Union	$IoU = \frac{tp}{tp+fp+fn}$	$I_{iou} = z \times \sqrt{\frac{IoU \times (1-IoU)}{N}}$	

TABLE 2.1 – Evaluation criteria of ROI detection methods. tp, fn, fp, tn are a number of pixels. $\beta = 0.3$, $N = 1156$.

where N is the number of testing samples. In this experiment, $N = 1156$ and $z = 1.96$ for 95% confidence interval.

2.2.4.2 Results and discussion

Examples of different ROI maps are shown in Fig. 2.11. Comparing the results in binary scale (see Fig. 2.11(b), (d), (f), (h), (j) and (k)), the results at rows (j) and (k) representing our ROIE methods are better since they are smoother, have more precise details and less background noise than other results. Tang's results do not seem precise in the case of the 2 first columns since their results mainly focus on few sharp details of the 2 close-up images. The results for large field images seem better than those of the close-up images. Aydin's results look better than Tang's results but they are still not good enough. The color saliency maps generated by Perazzi's method and Zheng's method at the two first columns of row (e) and row (g) are better than those of Aydin's and Tang's results but the results are not really good for large field images when sharpness factor is dominant. The main superiority of our methods is the high accuracy in both cases when photographers consider either sharpness or color saliency to define ROIs. The evaluations for the estimated sharpness maps, color saliency maps, handcrafted ROI maps and ROI maps generated by the deep models on the dataset are presented in Fig. 2.12, Fig. 2.13, Fig. 2.14, Fig. 2.15 and Fig. 2.16.

Firstly, the comparison between our sharpness maps and the two ROI maps based on sharpness information is shown in Fig. 2.12. Looking at the precision recall curves in Fig. 2.12, the AUC value of our sharpness maps is better than those of Aydin's maps (0.976 against 0.927). The column chart in Fig. 2.12 shows that the highest values of precision, recall, F-measure and IoU belong to our maps around 0.969 ± 0.010, 0.856 ± 0.005, 0.933 ± 0.014, 0.913 ± 0.016 respectively.

Secondly, Fig. 2.13 shows the comparison between our color saliency maps and the two

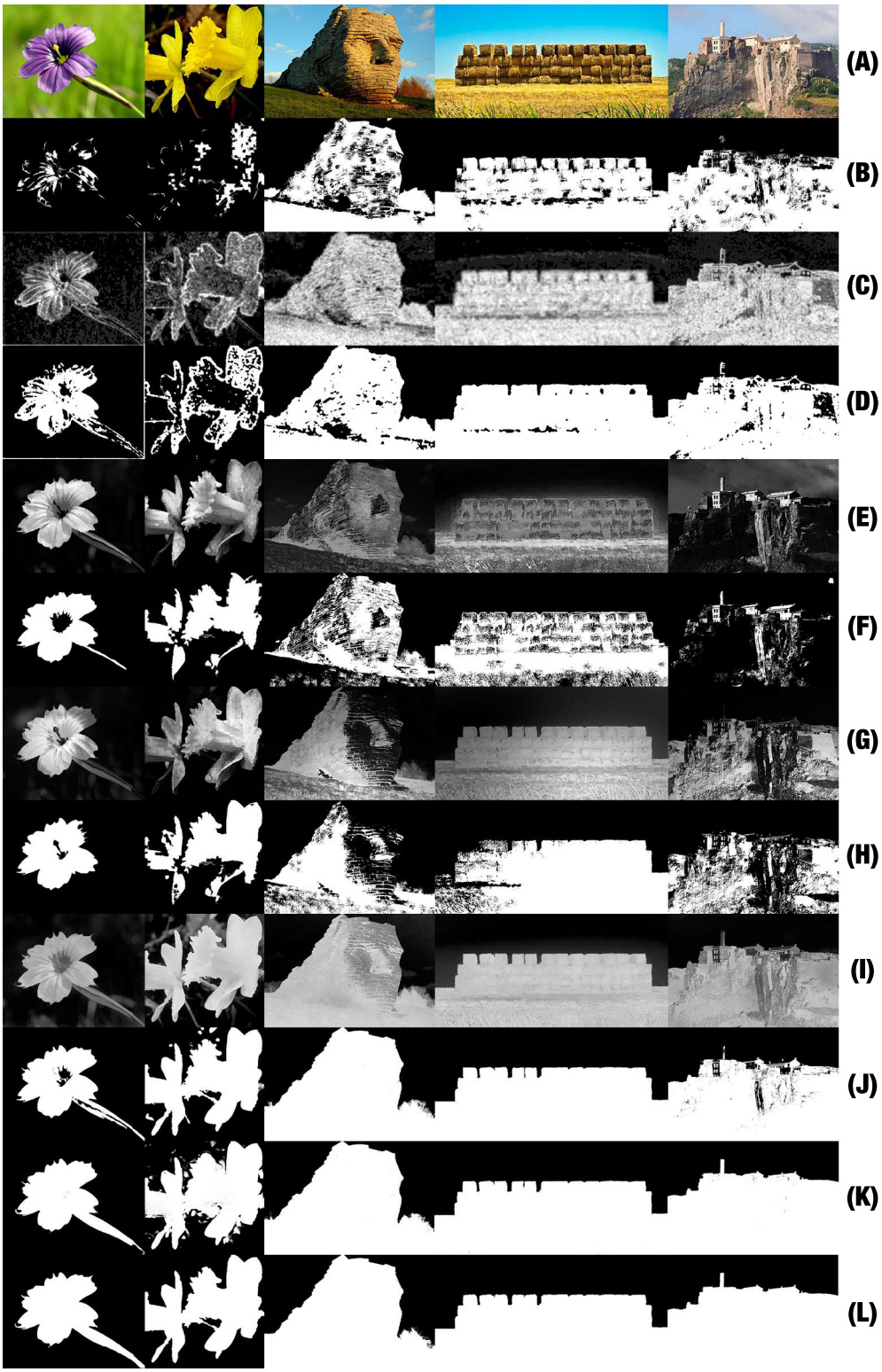


FIGURE 2.11 – Examples of ROI maps. (a) Original images. (b) Tang's [TLW13] sharpness maps. (c) and (d) Aydin's [ASG15] clearness maps and the binarized versions of them. (e) and (f) Perazzi's [Per+12] color saliency maps and the binarized versions of them. (g) and (h) Zheng's [ZZC13] color saliency maps and the binarized versions of them. (i) and (j) Handcrafted ROI maps based on both sharpness and color information and the binarized versions of them. (k) ROI maps generated by the first deep model. (l) ground truth.

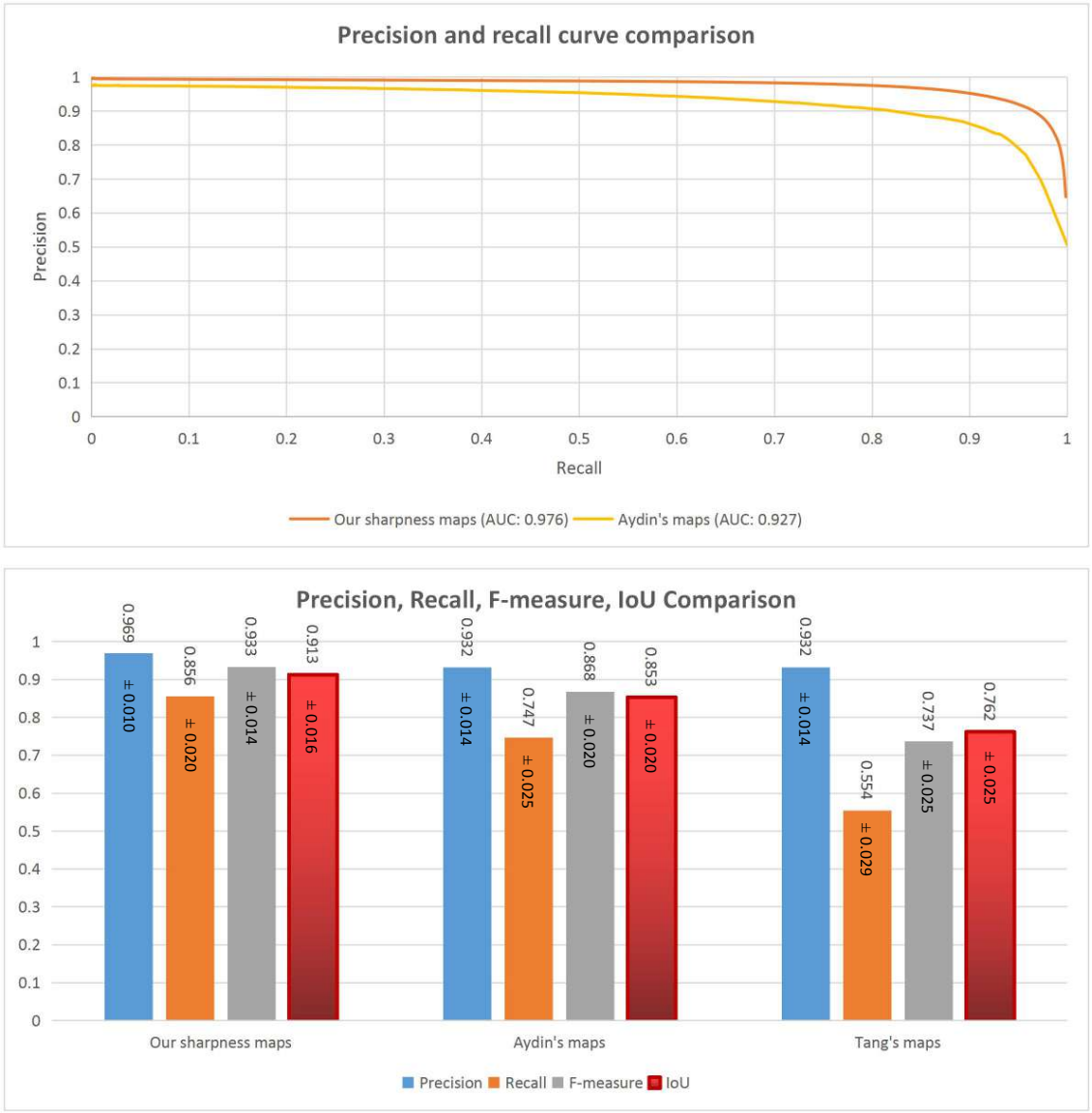


FIGURE 2.12 – Evaluations for the proposed sharpness maps, Aydin's maps and Tang's maps on the dataset. Tang's ROI results are binary maps so it is not necessary to consider their precision and recall curve and to apply a threshold on those maps.

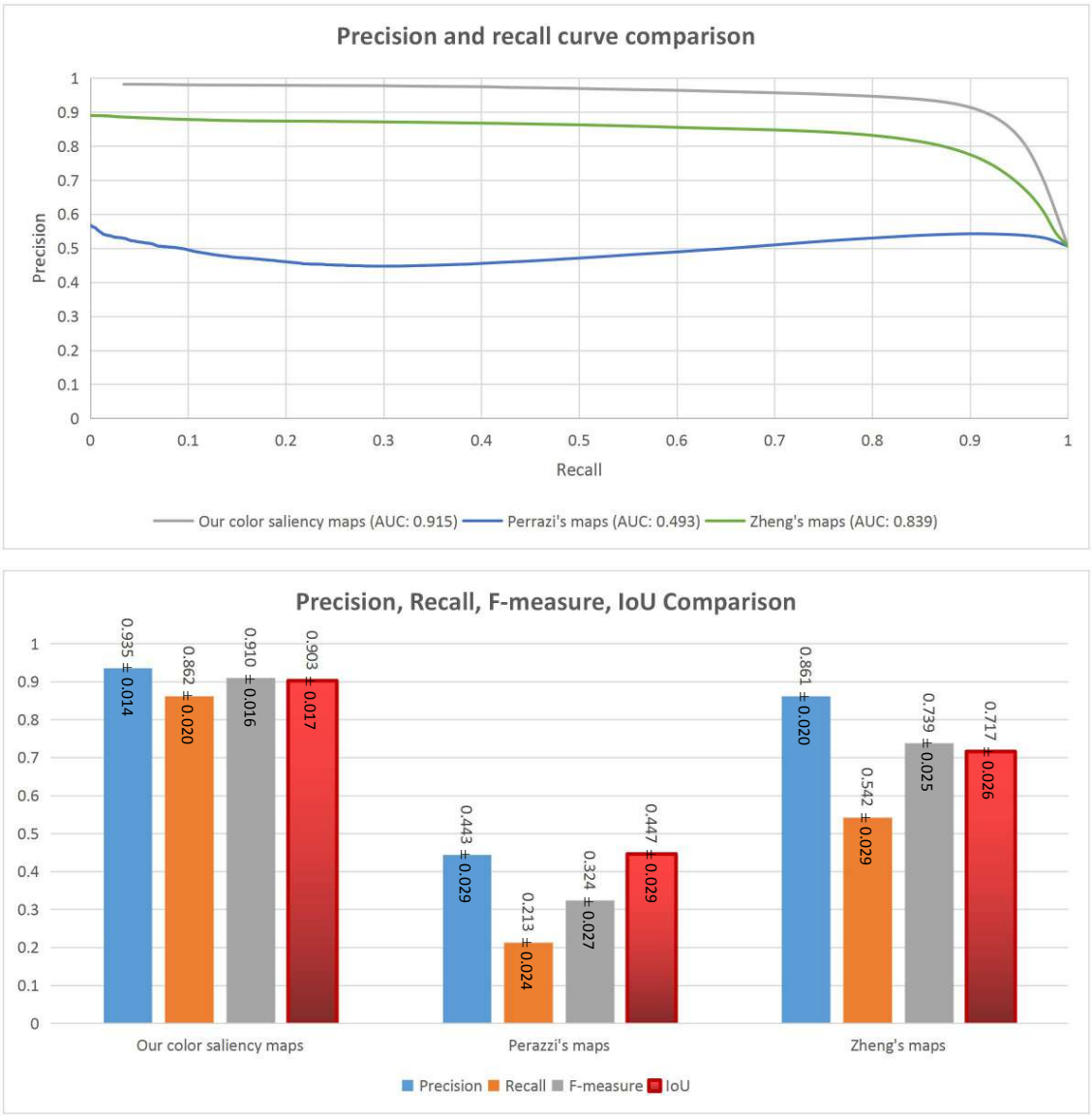


FIGURE 2.13 – Evaluations for the proposed color saliency maps, Perazzi's maps and Zheng's maps on the dataset.

ROI maps based on color contrast information. It appears that the highest values of AUC (0.915), precision (0.935 ± 0.014), recall (0.862 ± 0.020), F-measure (0.910 ± 0.016) and IoU (0.903 ± 0.017) are associated to our maps. The cause of the bad results of Perazzi's maps might be the differences between their color saliency definition and our color saliency definition since Perazzi et al. mostly focus on color contrast between all regions so regions having the most different colors are considered as the regions with the highest color saliency levels. About Zheng's method, they consider initially the colors of the center regions as salient so their results are better than Perazzi's results. Comparing with Zheng's maps, our color saliency maps have better results. The proposed method focuses on both color contrast and colors of in-focus and center regions. A region is considered as a high color saliency region if its colors are different from the colors in the out-of-focus regions and similar to the in-focus or center regions.

The third comparison is for our sharpness maps, our color saliency maps and our handcrafted ROI maps. Looking at the graphs in Fig. 2.14, the results of the ROI maps are better than those of the sharpness maps and the color saliency maps with the highest AUC (0.986), precision (0.979 ± 0.008), recall (0.933 ± 0.014), F-measure (0.966 ± 0.010) and IoU (0.958 ± 0.012) values. It proves the efficiency of combining sharpness and color information to extract ROIs.

The comparison between the ROI maps generated by the 3 deep models is presented in Fig. 2.15. Generally, all the 3 models have good performances. The first model with 3 main components (encoding, transformation and decoding components) using residual blocks has the highest performance around 0.966 ± 0.010, 0.974 ± 0.009, 0.966 ± 0.011 and 0.973 ± 0.009 for precision, recall, F-measure and IoU values respectively. It reflects that the architecture with 3 main components (encoding, transformation and decoding) is the best one among the considered architectures and residual blocks seem better than convolutional blocks in this case.

Comparing the handcrafted ROIE method and the deep learning based method, the precision and F-measure values of the 2 methods are almost the same but the deep model has higher recall values (0.974 ± 0.010) and a better balance between precision, recall and F-measure than those of the handcrafted ROI detection method. Generally, the 2 proposed methods have impressive results in which the results of the deep learning based method are slightly better than those of the handcrafted method at 0.973 ± 0.009 versus 0.958 ± 0.012 for IoU values.

2.2.5 Conclusions

In this work, we point out that sharpness only or color saliency only are not enough to precisely define ROIs (regions attracting viewers' eyes) while the combination of the two factors improves the performances. This ROIE task has been studied with both handcrafted and deep learning based approaches. They have been tested and compared with 4 other ROIE methods on a dataset containing 1156 images with the ROI ground truth. The gained results are quite good for both proposed methods but the results of deep learning based method are slightly better so the deep learning based ROIE method is going to be considered in the next chapter. ROIE is a preparation step before computing ROI features and background features from the corresponding regions. The influence of ROI features and background features in IAA

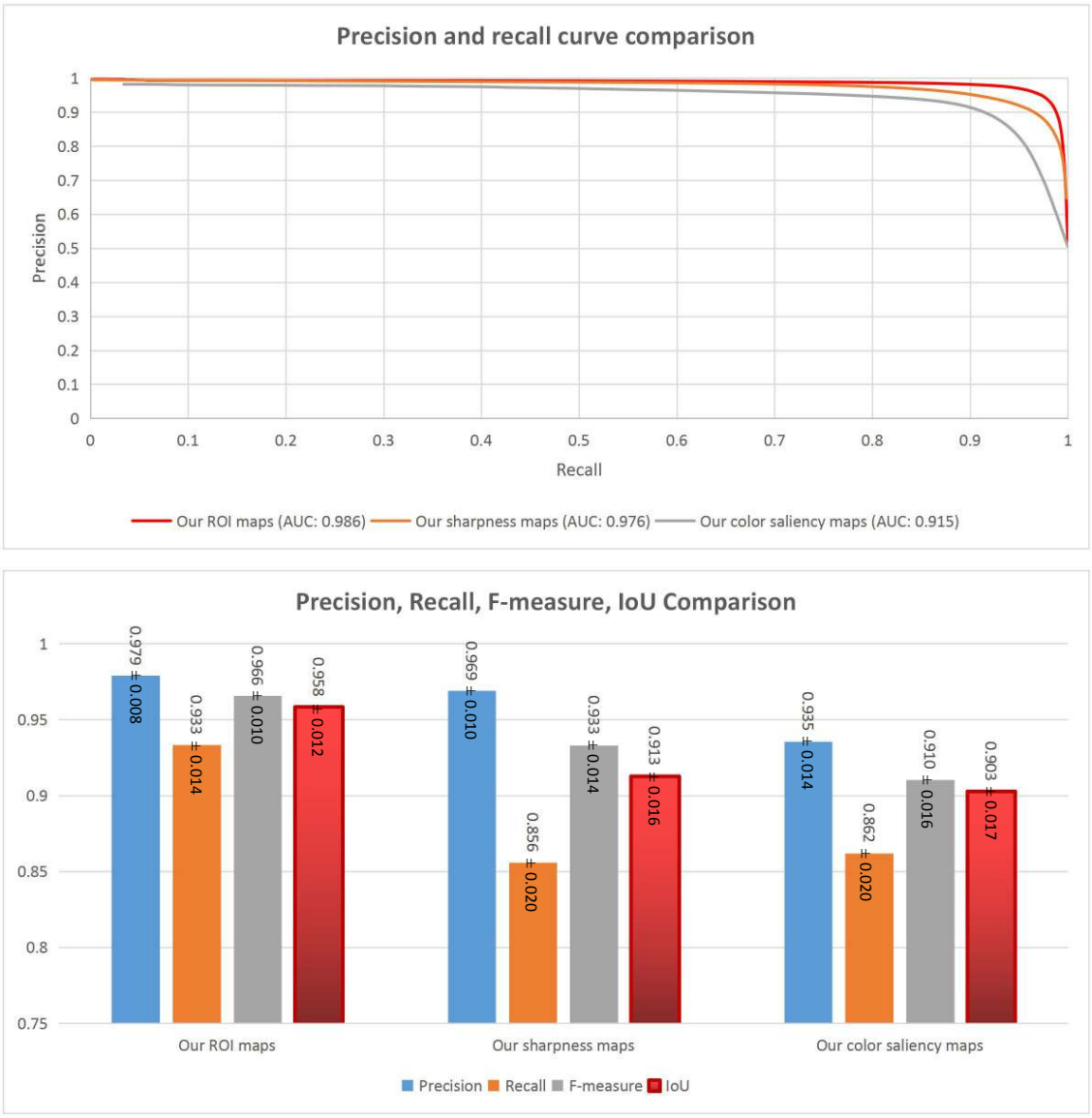


FIGURE 2.14 – Evaluations for the proposed ROI maps, the proposed sharpness maps and the proposed color saliency maps on the dataset.

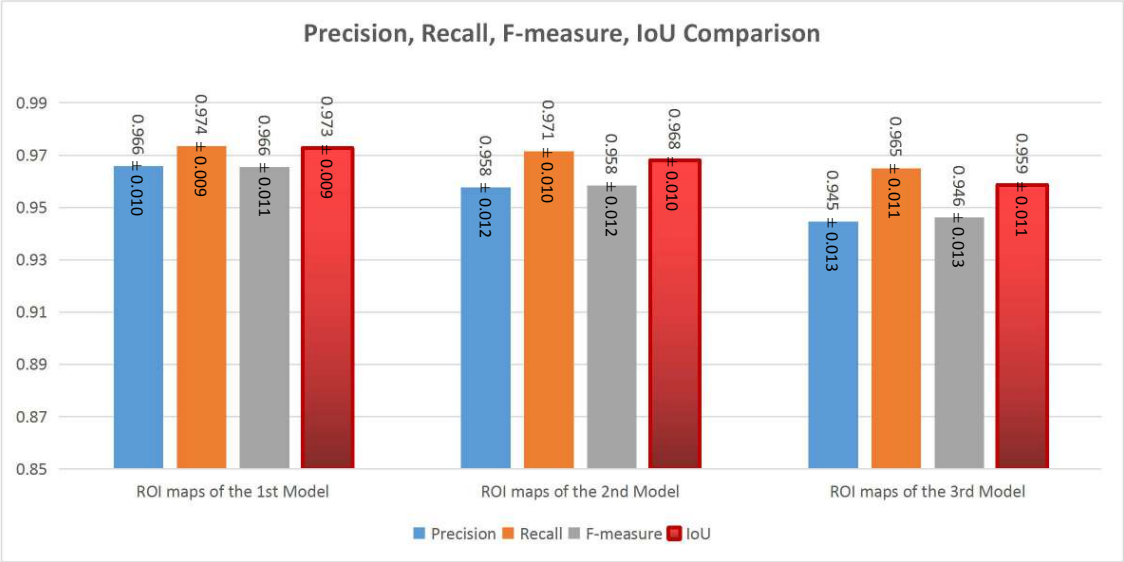


FIGURE 2.15 – Evaluations for the ROI maps generated by the 3 deep models on the dataset.

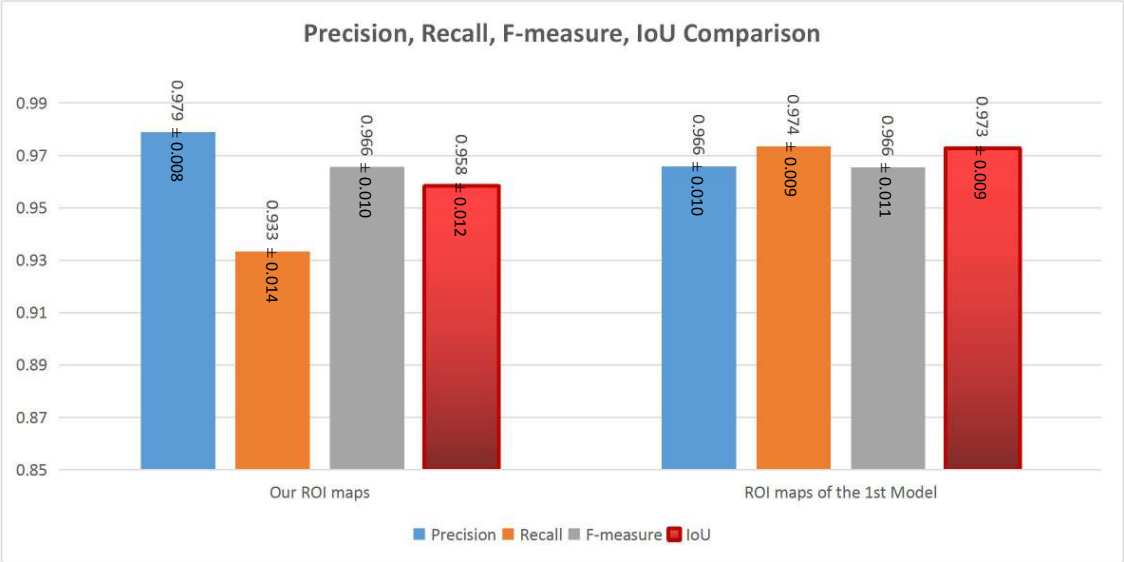


FIGURE 2.16 – Evaluations for the handcrafted ROI maps and the ROI maps generated by the deep model on the dataset.

is going to be estimated in the next chapter to answer the question "IAA : with or without prior region segmentation ?".

2.3 Large field / Close-up Image Classification (LCIC)

Large field images and close-up images are 2 typical image categories having opposite photographic rules for taking and assessing them (composition, brightness, contrast, distance,...) so the 2 categories have different criteria for IAA. Exploiting the classification between the 2 image categories could help improve the IAA performance so LCIC is studied in this section.

2.3.1 State of the art

Image classification has been studied for many years and the main idea is to use image features that are computed from image data either by hand [BZM07] ; [Ton+16] or via a learning algorithm [Guo+17] ; [He+18] to separate images into different categories. The focused problem in this chapter is to classify large field images and close-up images (image samples can be seen in Fig. 2.1). Until now, there are few researches about this classification. In [Wan], Wang et al. propose a method using color coherence vector and color moments to classify close-up and non close-up images. In another study, Zhuang et al. [Zhu+14] divide an image into 256 parts. The number of edge points in each part is counted to build a 256 bin histogram. The 256 bin values and standard deviation of those values are the key features to classify close-up and distance view images. In [Ton+16], Tong et al. use features representing the distributions of high frequencies in the first classification stage. In the second one, the spatial size and the conceptal size are used to classify distance / close-up view images. All features used in those classification methods are handcrafted features. The role of EXIF features and learned features for LCIC is still an open question.

Handcrafted features and learned features have been widely used for image classification [LW07]. Nowadays, deep learning approaches are the must for image classification [RW17]. At the same time, EXIF data has not been widely used for image classification. EXIF data are metadata (data information of data) and tags revealing photo information such as picture-taking time, picture-taking conditions [Tec02]. Surprisingly, EXIF features have been only occasionally used in researches. In [HCC08], Huang et al. use the manufacturer, camera model, date and time stamp and some other EXIF parameters as watermark information to protect image copyright. In [LF10], aperture, exposure value, ISO and picture-taking time are exploited to enhance ROI detection. In [BL04] ; [BL05], Boutell et al. integrate image content and EXIF data consisting of exposure time, flash use and focal length to classify in-door and out-door images.

In this section, the performances of LCIC based on EXIF features, handcrafted features or learned features are compared in terms of accuracy and computational complexity.

2.3.2 EXIF features for LCIC

In photography, camera tunnings are stored by digital cameras as EXIF data. 4 EXIF parameters and a combination of some of them are considered in this study.

2.3.2.1 Aperture

Aperture refers to the size of lens opening for light when a picture is captured. This parameter is stored as a *f-stops* value such as $f/1.4$, $f/2$, $f/2.8$,... in which $f\text{-}stops = \frac{f}{D}$ where f is the focal length and D is the diameter of the entrance in a camera. A smaller *f-stops* value represents a wider aperture. The Depth Of Field (DOF) and brightness of pictures are affected by the setting of aperture. See examples in Fig. 2.17 and Fig. 2.18, a decrease of the aperture value makes an increase of DOF and a decrease of brightness.

FIGURE 2.17 – Photos taken with different aperture settings. The left picture having low DOF is captured with a large aperture while the right picture having deep DOF is taken with a small aperture (image source : https ://photographylife.com).

2.3.2.2 Focal length

Focal length exhibits the distance from the middle of the lens to the digital sensor and it also decides the angle of view in the photo. This parameter is measured in millimeters. A long focal length makes a narrow view and a wide scene is captured with a short focal length (see Fig. 2.19).

2.3.2.3 Exposure time

Exposure time represents the total time for light falling on the sensor of a camera during shooting. It is measured in seconds. In weak light conditions (see Fig. 2.20(a)) or to create some special effects (see Fig. 2.20(b)), photographers use long exposure time. A short exposure time is regularly used when capturing moving objects like taking sport photos (see Fig. 2.20(c)).

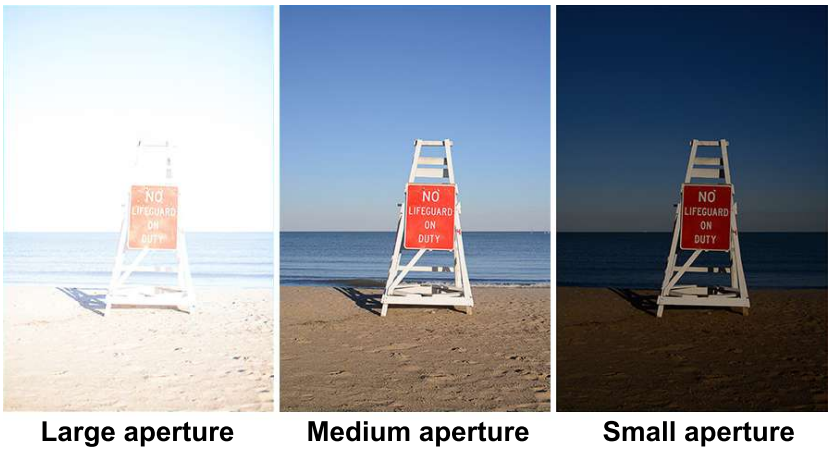


FIGURE 2.18 – Influence of aperture on photo brightness. The wider aperture is used, the brighter picture is taken (image source : https ://photographylife.com).

FIGURE 2.19 – Examples of photos taken with different focal lengths (image source : https ://www.colesclassroom.com).

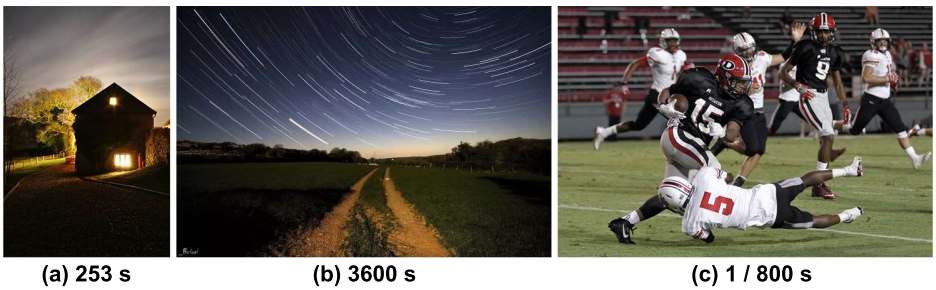


FIGURE 2.20 – Examples of pictures are taken with different exposure time for different purposes (image source : https ://photo.stackexchange.com and https ://digital-photography-school.com).

2.3.2.4 ISO

ISO describes the sensitivity level of the sensor in a camera. ISO parameter is measured with numbers such as 100, 200, 400,... The lower ISO value represents the less sensitive mode of the sensor. The brightness of a photo decreases with the decrease of ISO (see Fig. 2.21). However using a too sensitive mode could generate some noise in the taken photo.

FIGURE 2.21 – Pictures taken with different ISO modes. When increasing ISO, the camera sensor is more sensitive to light and the photo looks brighter (image source :https ://photographylife.com).

2.3.2.5 Illumination measure

Illumination measure refers to the light falling on a surface [HS11]. This feature is calculated as :

$$I_m = \log_{10}(\frac{aperture^2}{exposure\ time}) + \log_{10}(\frac{250}{ISO}) \tag{2.23}$$

2.3.2.6 EXIF feature selection

In this part, the influences of EXIF features on LCIC are investigated. At the first step, EXIF values of 400 large field and 400 close-up photos (the training set in the next LCIC experiments) coming from the Flickr dataset are displayed Fig. 2.22. It appears that the differences of EXIF parameters between close-up and large field images are significant in aperture, focal length, illumination measure and to a smaller extent in exposure time.

Unsurprisingly the aperture data is very efficient to distinguish between close up and large field images. Actually, a high aperture value is regularly chosen to highlight the objects by low DOF effect. In the other hand, because large field scenes are far from the camera, a small

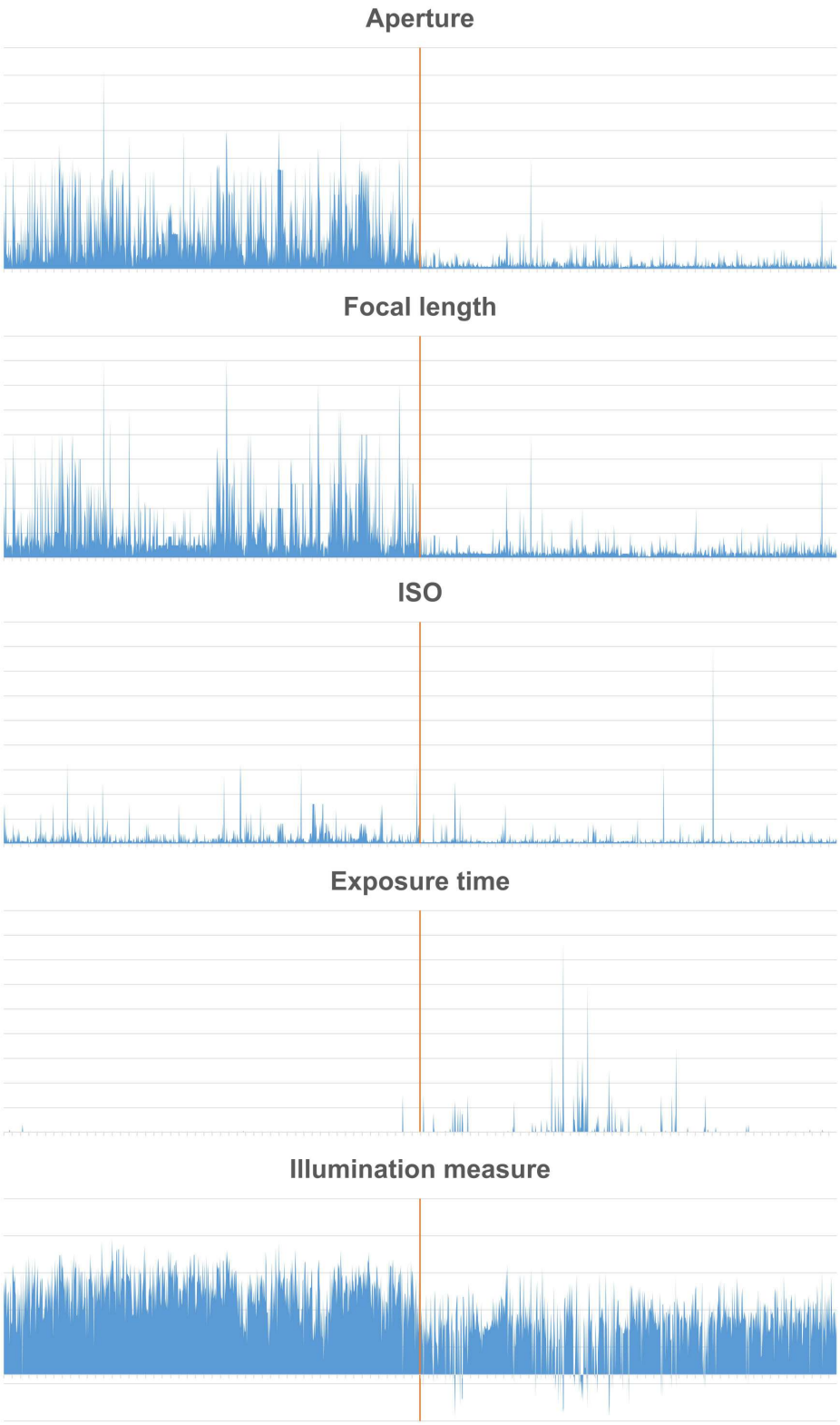


FIGURE 2.22 – The distribution of EXIF values on 400 close-up images (the left side) and 400 large field images (the right side).

aperture setting is set for capturing a large field photo to gain a high DOF. Focal length is the second discriminating feature. A large field scene is wide so photographers often use a short focal length to get the whole scene. In contrast, to focus on close-up objects, a longer focal length is regularly chosen to take close-up photos. Illumination measure and exposure time are also going to be considered for LCIC. On the contrary, ISO feature is not relevant enough.

2.3.3 Handcrafted features for LCIC

The main goal of this part is to build a handcrafted feature set for LCIC based on usual features computed from image data. Firstly, a large handcrafted feature set is built from common handcrafted features appearing in different researches [Vai+99] ; [Dat+06] ; [KTJ06] ; [LT08] ; [ASG15]. The initial handcrafted feature set includes 2030 features related to hue, saturation, brightness, red, green and blue channels, sharpness, color saliency and contrast. Those features are global features (features computed from the whole image) and local features (features computed for different local regions). The local features are computed from ROIs, background and regions split by symmetry rules, landscape rule, rule of thirds (see Fig. 2.23).

FIGURE 2.23 – Illustrations of region splits. The first row shows the whole scene and regions split by landscape rule and rule of thirds respectively. The second row presents regions split by symmetry rules.

A feature reduction algorithm is applied on the feature set to select the most relevant features related to the task. To perform the algorithm, the relevance of each feature needs to be evaluated. This is done by using the relief method [KR92]. The evaluation set S containing 2 subsets S_{C1} and S_{C2} (corresponding to the 2 image categories) is considered to compute the relevance of the features. All features of each image in S are calculated and normalized to

range [0,... 1]. The relevance of a given feature f is calculated as :

$$r(f) = dif(f, S_{C1}, S_{C2}) - dif(f, S_{C1}, S_{C1}) - dif(f, S_{C2}, S_{C2}) \tag{2.24}$$

$$dif(f, X, Y) = \frac{\sum_{i=1}^{\| X \|} \sum_{j=1}^{\| Y \|} (d(f, X_i, Y_j))}{\| X \| \times \| Y \|} \tag{2.25}$$

where the number of images in set X is presented as $\| X \|$. X_i is the i^{th} image of the set X while the absolute difference between f values of the 2 images x and y is represented as $d(f, x, y)$. The feature relevances are then normalized to range [0,... 1]. The highest $r(f)$ values illustrate the most relevant features.

In order to reduce the number of features and keep the most relevant features F_T, it is necessary to find a threshold T to be applied on feature relevance R to discard irrelevant features (features having the relevance smaller than the threshold T). To find the threshold, an algorithm based on the feature relevance and the binary search algorithm is applied [LE+19](cf. Fig. 2.24). The first step of the algorithm is to initialize a lower threshold T_1 and an upper threshold T_2 to 0 and 1 respectively. T_1 and T_2 are then considered as the thresholds to select 2 feature sets F_{T_1} and F_{T_2} ($F_{T_j} = \{f_x | r_x \geq T_j\}$). F_{T_1} and F_{T_2} are then applied to classify large field and close-up images by using 2 Support Vector Machine (SVM) models. Comparing the 2 models trained on S_1 (containing 50% of S_{C1} and 50% of S_{C2}) and tested on S_2 ($S = S_1 \cup S_2$) can point out which threshold is the best (T_1 or T_2). The best threshold is kept while the worst threshold is updated to reduce the distance between the 2 thresholds. After performing K iterations, the final threshold T is computed as the average of the 2 thresholds T_1 and T_2. The algorithm is then applied on the feature sets to keep the most relevant features only.

After the most relevant features are selected, there is an additional step for handcrafted features only : Analyzing selected features to understand them in order to remove overlapping features and to optimize features (it is not applied for learned features because learned features are not easy to understand). To analyze a selected feature, the first step is to build graphs for the distribution of feature values in the 2 categories. For example, 9 selected features for LCIC are analyzed in Fig. 2.25. Those features are means of gradient values in 9 regions R_1, R_2,... R_9 split by the rule of thirds. In close-up images, the means of gradient values in R_1, R_2 and R_3 are higher than those of large field images in which the difference in R_2 is the most significant. In contrast, in R_7, R_8 and R_9, large field images have higher gradient values than those of close-up images. In general, the mean of gradient values in R_4 and R_6 of close-up images are smaller than those of large field images and the mean of gradient values in R_5 of close-up images is higher than that of large field images but those differences are not significant.

Among selected features, there might be some overlapping features. For examples, sharpness information is presented in various ways : number of edge pixels (those pixels are determined by applying a mean threshold on the gradient map), mean, standard deviation, kurtosis, skewness of gradient values. Considering the differences of the feature values distributed in the 2 image categories, some of them could be removed and only features having the most significant differences are kept to simplify the feature set. Beside that, some local features are also overlapping. For instance, when considering gradient values in the 3 regions R_7, R_8,

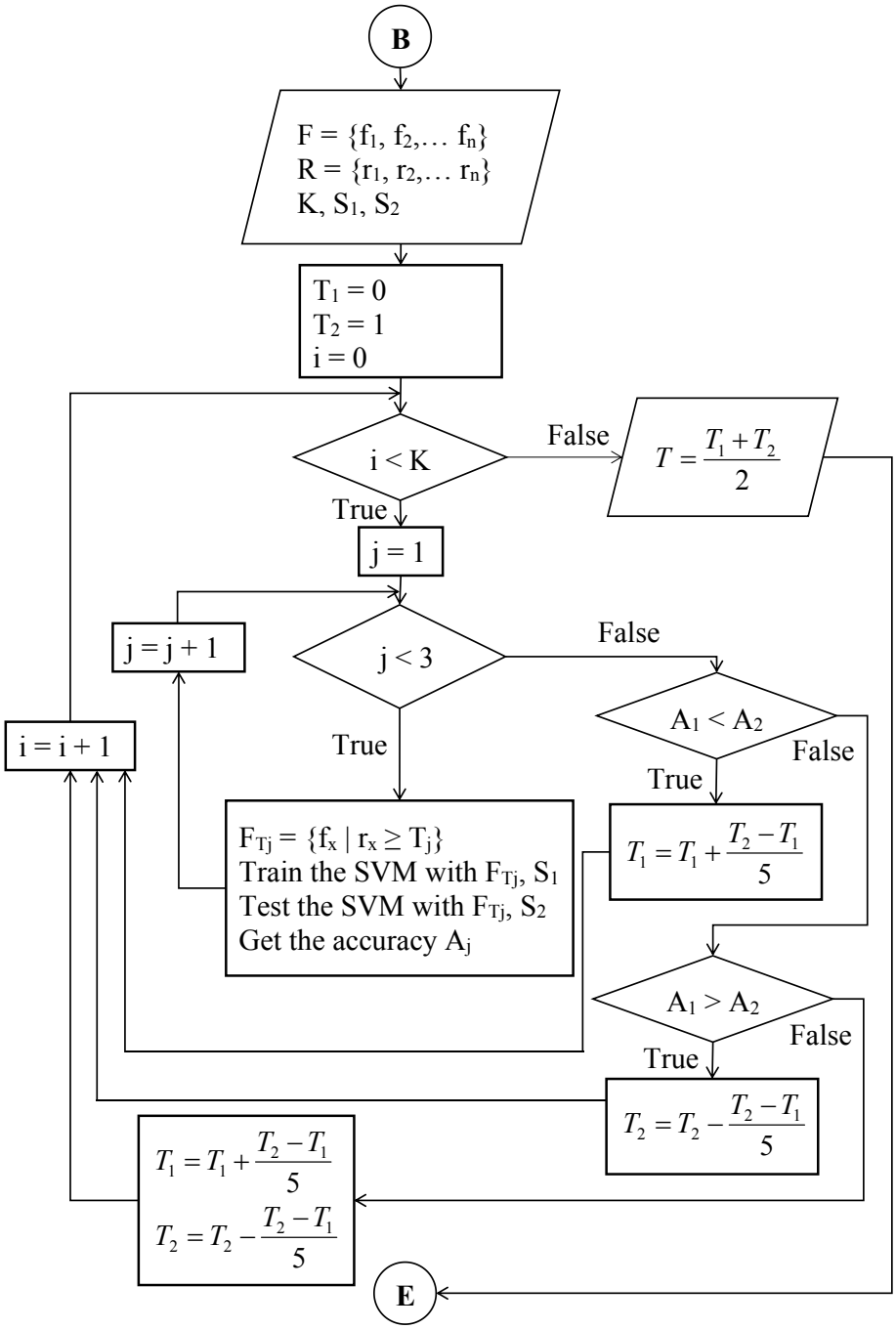


FIGURE 2.24 – Flowchart of the algorithm finding the optimal threshold. The inputs include the feature set F, the feature relevance set R, the number of iterations K, the training set S_1 and the testing set S_2. T_1, T_2 are the lower and upper thresholds respectively. F_{Tj} is the reduced feature set selected with the threshold T_j. A_j is the accuracy of the SVM classifier trained and tested with S_1 and S_2 respectively with the feature set F_{Tj}. The output of the algorithm is the optimal threshold T.

R_9, the values on the side of large field images are higher than those on close-up image side. It means that the gradient value of the region merged from R_7, R_8 and R_9 (that region is obtained by applying the landscape rule, see Fig. 2.23) of large field images is also higher than that of close-up images. Thus, the gradient value in the merged region is an overlapping feature in this case.

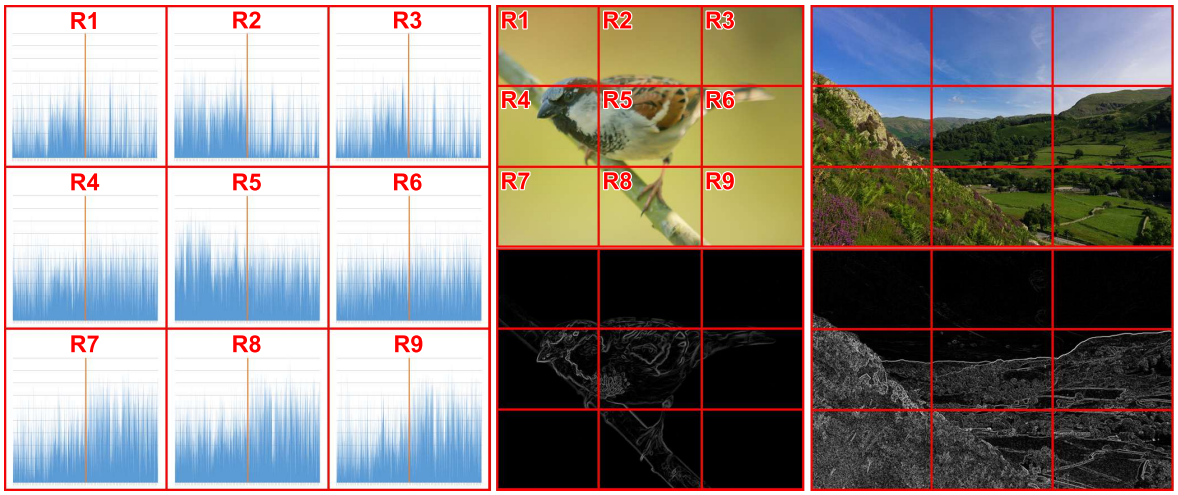


FIGURE 2.25 – Feature analysis. Left graphs : Distributions of mean of gradient values in 9 regions split by rule of thirds. The left side of each graph presents the distributions for close-up images while the right side presents the distributions for large field images. Right images : the first row contains examples of close-up images and large field images while the second row presents the corresponding gradient maps.

Running the algorithm on 1200 large field images and 1200 close-up images coming from the CUHKPQ dataset [TLW13] in which a half of them is used in the training phase (S_1) and the remaining is used in the testing phase (S_2), the 21 most relevant features are selected from 2030 features for the LCIC task (see overview of the features in Table 2.2).

2.3.4 Learned features for LCIC

Beside being handcrafted from images, features can also be learned by employing deep learning [KSH17]. VGG16 [SZ15] is a well-known deep CNN. It includes 3 main parts including convolutional layers, fully connected layers and a prediction layer. If the prediction layer is removed, that model can be considered as a feature extractor. From images of size 244×244, 4096 features are extracted by the VGG16 without the last layer. Although those features have been learned for the task of classifying objects in images, they can be applied for different tasks [PY10] such as IQA [JSS16] ; [TM18]. In this study, the VGG16 without the prediction layer pre-trained on the ImageNet dataset for the task of classifying objects in images is considered to compute the learned features for LCIC on the corresponding dataset. Instead of transferring all learned features, the most relevant features are selected because some of

Features	**Formula**
Sharpness features	$f_1 = \frac{1}{\|R_2\|} \times \sum_{(x,y)\in R_2} G(x,y)$ $f_2 = \frac{1}{\|R_7\|} \times \sum_{(x,y)\in R_7} G(x,y)$ $f_3 = \frac{1}{\|R_9\|} \times \sum_{(x,y)\in R_9} G(x,y)$ $f_4 = \sqrt{\frac{1}{\|R_5\|} \times \sum_{(x,y)\in R_5} (G(x,y) - \mu_g^{R_5})^2}$ $f_5 = \frac{\mu_g^{R_1} - \mu_g^{R_7}}{\mu_g^{R_1} + \mu_g^{R_7}}$ $f_6 = \frac{\mu_g^{R_2} - \mu_g^{R_8}}{\mu_g^{R_2} + \mu_g^{R_8}}$ $f_7 = \frac{\mu_g^{R_3} - \mu_g^{R_9}}{\mu_g^{R_3} + \mu_g^{R_9}}$ $f_8 = \sqrt{\frac{1}{w \times h} \times \sum_{x=1}^{w} \sum_{y=1}^{h} (G(x,y) - \mu_g)^2}$ $G(x,y)$ is the gradient value at point (x,y) $\| X \|$ is the number of pixel in region X $\mu_g^{R_i}$ is the mean of gradient values in region R_i.
Color features	$f_9 = \frac{\mu_b^{R_1} - \mu_b^{R_7}}{\mu_b^{R_1} + \mu_b^{R_7}}$ $f_{10} = \frac{\mu_b^{R_2} - \mu_b^{R_8}}{\mu_b^{R_2} + \mu_b^{R_8}}$ $f_{11} = \frac{\mu_b^{R_3} - \mu_b^{R_9}}{\mu_b^{R_3} + \mu_b^{R_9}}$ $f_{12} = \sqrt{(\mu_{re}^{R_1} - \mu_{re}^{R_7})^2 + (\mu_{gr}^{R_1} - \mu_{gr}^{R_7})^2 + (\mu_{bl}^{R_1} - \mu_{bl}^{R_7})^2}$ $f_{13} = \sqrt{(\mu_{re}^{R_2} - \mu_{re}^{R_8})^2 + (\mu_{gr}^{R_2} - \mu_{gr}^{R_8})^2 + (\mu_{bl}^{R_2} - \mu_{bl}^{R_8})^2}$ $f_{14} = \sqrt{(\mu_{re}^{R_3} - \mu_{re}^{R_9})^2 + (\mu_{gr}^{R_3} - \mu_{gr}^{R_9})^2 + (\mu_{bl}^{R_3} - \mu_{bl}^{R_9})^2}$ $\mu_b^{R_i}$, $\mu_{re}^{R_i}$, $\mu_{gr}^{R_i}$ and $\mu_{bl}^{R_i}$ are the means of brightness, red, green and blue intensities in region R_i.
ROI / background features	$f_{15} = \frac{1}{\|R_2\|} \times \sum_{(x,y)\in R_2} (M_b^{roi}(x,y) > 0)$ $f_{16} = \frac{1}{\|R_7\|} \times \sum_{(x,y)\in R_7} (M_b^{roi}(x,y) > 0)$ $f_{17} = \frac{1}{\|R_9\|} \times \sum_{(x,y)\in R_9} (M_b^{roi}(x,y) > 0)$ $f_{18} = \frac{1}{\|ROI\|} \sum_{x=1}^{w} \sum_{y=1}^{h} G^{roi}(x,y)$ $G^{roi} = M_b^{roi} \times G$ $f_{19} = \frac{\|R_{M^{roi}}\|}{\|R_{\neg M^{roi}}\|}$ $f_{20} = \frac{\|R_{M^{roi}} \cap R_{\neg M^{roi}}\|}{\|R_{M^{roi}} \cup R_{\neg M^{roi}}\|}$ $f_{21} = \frac{\mu_b^{roi} - \mu_b^{bg}}{\mu_b^{roi} + \mu_b^{bg}}$ $\| ROI \|$ is the number of pixels in the ROIs G is the gradient image M^{roi} is the ROI map, R_I is the distribution rectangle of pixel values in image I and $\neg I$ is the inverted image of I

TABLE 2.2 – Overview of the proposed handcrafted features for LCIC.

them are pre-learned for a different task so they could not be relevant for the LCIC task so the feature reduction algorithm described before is run on 1200 large field images and 1200 close-up images coming from the CUHKPQ dataset to select the 925 most relevant features from the 4096 features learned by the VGG16 (see Fig. 2.26).

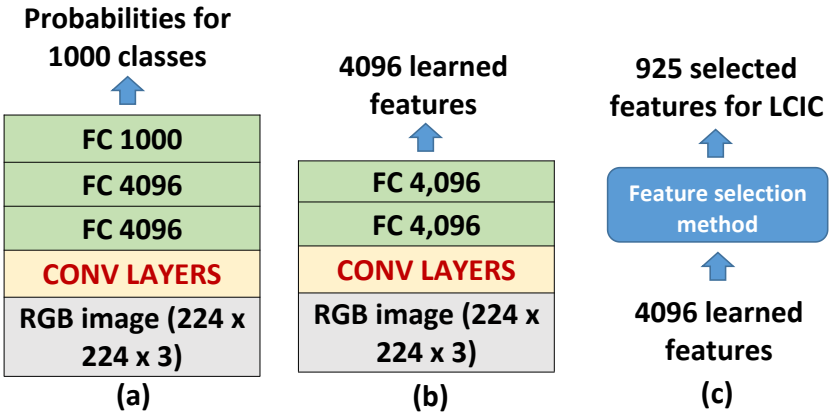


FIGURE 2.26 – The feature selection process among the features learned by VGG16. From left to right : (a) The structure of the VGG16 pre-trained on ImageNet dataset for the purpose of classifying images into 1000 classes. (b) The structure of the feature extractor based on the pre-trained VGG16. (c) The process to select the 925 most relevant features to perform LCIC.

2.3.5 Experiment and results

2.3.5.1 Dataset and setup

LCIC is performed separately with EXIF, handcrafted and learned features. In order to evaluate the influences of the different feature types fairly, an SVM classifier is trained and tested to evaluate the classification performances obtained with each feature set because of its simplicity. If complex classifiers had been used, the accuracy of the classifications could be affected not only by the input features but also by the suitability between the model structure and input features. The experiments are performed on 1600 images (with EXIF data) including 800 large field and 800 close-up images collected and categorized from Flickr website by the authors. Half of the large field and close-up images are selected randomly to train the classifiers while the others are used to test. Each SVM classifier is applied with $C = 0.5, g = 10^{-5}, e = 1.192 \times 10^{-7}$ and different kernels : Poly, Linear, RBF and Sigmoid to find the most appropriate kernel. After performing the experiment only the best results with a Linear kernel are presented.

LCIC is evaluated based on the Accuracy (A) depending on TP, TN, FP, FN (true positive, true negative, false positive and false negative expressed as a number of images) described in Table 2.3.

Evaluation criteria	**Formula**
Accuracy	$A = \frac{TP+TN}{TP+FP+TN+FN}$
Confidence interval of accuracy	I_a
Lower accuracy	$A_l = A - I_a$
Upper accuracy	$A_u = A + I_a$
Feature computational time	T_F
Classification time	T_C
Total computational time	$T_T = T_F + T_C$

TABLE 2.3 – Overview of evaluation criteria for LCIC.

LCIC using the 4 EXIF features			
		Prediction	
		Close-up image	Large field image
Ground truth	Close-up image	TP = 348	FN = 52
	Large field image	FP = 46	TN = 354
A = 0.878	I_a = 0.023	A_l = 0.855	A_u = 0.901
T_F = 1 ms	T_C = 1 ms	T_T = 2 ms	

TABLE 2.4 – LCIC using the 4 EXIF features.

The experiments have been conducted on a PC equipped with an Intel Core i7-2670QM CPU 2.40 GHz and 11.9 GB memory to evaluate the feature computational time T_F (the time for computing features from images directly) and the classification time T_C (the time for classifying images based on computed features) and the total computational time ($T_T = T_F + T_C$) per image. Additionally, the computational time for learned features is often smaller if there they are computed with an GPU so an GPU NVIDIA Quadro P400 is used to compute the learned features (the computational time for handcrafted, EXIF features in this experiment are not affected by the GPU).

2.3.5.2 Results and discussions

EXIF features based LCIC : The results of LCIC using the 4 EXIF features are presented in Table 2.4. Using a very small number of simple features (only 4 features), the classification accuracy at 0.878 ± 0.023 is impressive. Additionally, the feature computational time for EXIF features is very small (under 1 ms because there is only one simple EXIF feature that needs to be computed).

Handcrafted features based LCIC : Table 2.5 shows the results of the classification using the proposed handcrafted features. The handcrafted feature set is simple since it includes only 21 features but its reference classification rate is also impressive (the overall accuracy is

LCIC using the 21 handcrafted features			
		Prediction	
		Close-up image	Large field image
Ground truth	Close-up image	TP = 349	FN = 51
	Large field image	FP = 51	TN = 349
A =ca 0.873	$I_a = 0.023$	$A_l = 0.850$	$A_u = 0.896$
$T_F = 30$ ms	$T_C = 1$ ms	$T_T = 31$ ms	
LCIC using Wang's feature set (105 features)			
A = 0.774	$I_a = 0.029$	$A_l = 0.745$	$A_u = 0.803$
LCIC using Zhuang's feature set (257 features)			
A = 0.854	$I_a = 0.024$	$A_l = 0.830$	$A_u = 0.878$

TABLE 2.5 – LCIC using the 21 handcrafted features compared with LCIC using other handcrafted feature sets.

0.873 ± 0.023). In order to prove the efficiency of our handcrafted features, the classification based on those features is compared with the classifications based on other handcrafted features including Wang's [Wan], Zhuang's [Zhu+14] features. Despite of using more features, the classifications with Wang's (105 features) and Zhuang's (257 features) feature sets have lower accuracy at 0.774 ± 0.023 and 0.854 ± 0.024 respectively. Those results prove the efficiency of our handcrafted features.

Learned features based LCIC : The results of classification with the 925 most relevant features learned from the VGG16 are shown in Table 2.6. Obviously, the classification with learned features has the highest overall accuracy (0.989 ± 0.007) but the number of features is also the biggest (925 features) and the feature computational time is also the longest (434 ms - without the GPU) among the studied feature sets. With the GPU, the computational time is much smaller (16 ms).

Comparisons : To start with, it appears that EXIF features are quite powerful for LCIC since the accuracy at 0.878 ± 0.023 is obtained with only 4 EXIF features. With handcrafted features, the number of features is higher (21 versus 4) while the classification accuracy is almost the same (0.873 ± 0.023). Secondly, the classification with learned features has the highest accuracy (0.989 ± 0.007). However the number of learned features is also the biggest (925 learned features against 21 handcrafted features and 4 EXIF features).

In order to compare the role of those features accurately, the classifications using the top 21 and top 4 most relevant learned features are performed and the results are shown in Table 2.6. The comparisons between LCIC using the reduced VGG16 feature sets and LCIC using the handcrafted features and EXIF features are presented in Table 2.7. It appears that the learned features are very efficient for LCIC since with the same number of features as handcrafted features (21 features) the accuracy of the classification based on the 21 most relevant learned

LCIC using the 925 most relevant VGG16 features			
		Prediction	
		Close-up image	Large field image
Ground truth	Close-up image	TP = 392	FN = 8
	Large field image	FP = 1	TN = 399
A = 0.989	$I_a = 0.007$	$A_l = 0.982$	$A_u = 0.996$
Without the GPU	$T_F = 434$ ms	$T_C = 2$ ms	$T_T = 436$ ms
With the GPU	$T_F = 16$ ms	$T_C = 2$ ms	$T_T = 18$ ms
LCIC using the 21 most relevant VGG16 features			
A = 0.981	$I_a = 0.009$	$A_l = 0.972$	$A_u = 0.990$
Without the GPU	$T_F = 434$ ms	$T_C = 1$ ms	$T_T = 435$ ms
With the GPU	$T_F = 16$ ms	$T_C = 1$ ms	$T_T = 17$ ms
LCIC using the 4 most relevant VGG16 features			
A = 0.975	$I_a = 0.011$	$A_l = 0.964$	$A_u = 0.986$
Without the GPU	$T_F = 434$ ms	$T_C = 1$ ms	$T_T = 435$ ms
With the GPU	$T_F = 16$ ms	$T_C = 1$ ms	$T_T = 17$ ms

TABLE 2.6 – LCIC using the top 925, top 21 and top 4 most relevant learned features.

Feature set	$A \pm I_a$	T_F (ms)	T_C (ms)	T_T (ms)
EXIF features	0.878 ± 0.023	1	1	2
Top 4 most relevant learned features	0.975 ± 0.011	16	1	17
Handcrafted features	0.873 ± 0.023	30	1	31
Top 21 most relevant learned features	0.981 ± 0.009	16	1	17
Top 925 most relevant learned features	0.989 ± 0.007	16	2	18

TABLE 2.7 – LCIC based on the 4 EXIF features, 21 handcrafted features, top 925, top 21 and top 4 most relevant learned features.

features is higher than that of the handcrafted features (0.981 ± 0.009 versus 0.873 ± 0.023). Similarly, with only 4 learned features as EXIF features, the accuracy of the classification based on the 4 most relevant learned features is 0.975 ± 0.011, a very high accuracy while the classification accuracy with EXIF features is smaller (0.878 ± 0.023).

Fig. 2.27 shows the top 9 best classifications (images being classified correctly and having the biggest distances to the hyper-plane of the SVM classifiers) and the top 9 worst classifications (images being classified incorrectly and having the biggest distances to the hyper-plane) of each category. It appears that the feature sets are acting totally differently since there are no overlapping images between those results. The best classified close-up images using EXIF features are mostly low DOF images because of wide aperture values. Almost all the best close-up images (7 of 9) have high apertures, high illumination measures and long exposure time ($aperture \geq 10$ *and* $I_m \geq 4.0$ *and exposure time* $\geq \frac{1}{250}$) while no image of the best or worst large field photos and only one of the worst close-up images satisfies this condition. Additionally, 6 of the 9 best large field images have small focal lengths ($focal\ length \leq 50$), short exposure time ($exposure\ time \leq \frac{1}{250}$) and illumination measures ranging from 2.75 to 3.418 while no image of the best close-up photos and only one of the worst large field images have EXIF data in those ranges.

With handcrafted features the best classified close up images almost have blank background because some features are handcrafted to estimate the number of background details of close-up images (those features cannot be used to classify blank background or blur background) so the classifier focuses on blank background.

Because VGG16 have been pre-trained on the ImageNet dataset for the purpose of classifying objects in images, the extracted features have been designed to recognize objects very well. It explains why the top classified close-up images using those features are images with fish, bird, chicken, insect. Additionally, learned features seem to focus on the high frequency details in foreground of close-up images. In contrast, the differences between the best large field image classifications and the differences between the worst classifications are not clear.

Last but not least, the feature computational time and classification time per image are shown in Table 2.4. It is clear that EXIF features are the simplest ones when only one EXIF feature (illumination measure) needs to be computed and its feature computational time is only 1 ms. In contrast, without the GPU, the feature computational time of learned features is over 14 times of the handcrafted features (434 ms versus 30 ms). Additionally, the feature computational costs for the 21, 925 or 4096 learned features are the same because the feature extractor always computed all 4096 features. With the GPU, the computational time of the learned features decreases significantly to 16 ms (approximately 50% of the computational time of the handcrafted features). Although the time of SVM classification based on the computed features is almost the same (1 to 2 ms), the differences in the total classification time between those feature sets are significant. It points out that the classification based on EXIF features is very fast (only 2 ms). The classification based on handcrafted features is slower (30 ms) while without the GPU, the classification with learned features is very slow (434 ms) but the accuracy is not increasing in the same proportions. However, with the GPU, the weakness of the computational time for learned features is solved.

FIGURE 2.27 – The best and the worst classification based on different feature types .The first, second, third and fourth rows (separated by the red lines) present the best close-up, large field image classifications (images being classified correctly and having the biggest distances to the hyper-plane) and the worst large field and close-up image classifications (images being classified incorrectly and having the biggest distances to the hyper-plane) based on the EXIF, handcrafted and learned features respectively. A : Aperture, F : Focal length. E : Exposure time, I : Illumination measure.

2.3.6 Conclusions

In this part, 3 types of features including handcrafted features, learned features and EXIF features have been studied for LCIC. Their performances are evaluated in terms of classification accuracy, complexity, running time. It appears that learned features are very powerful for that task although they are complex, they require a strong GPU to reduce the computational time and it is not easy to understand them. EXIF features are quite efficient for LCIC since it is possible to obtain the same and quite good classification score by using 4 very simple EXIF features than by using 21 complex handcrafted features. EXIF features are simple, efficient but unfortunately they are not always available.

2.4 Conclusions

ROIE and LCIC are the preparation steps before performing IAA in the next chapter. Firstly, starting with the results of LCIC, IAA based on those results are studied and it is then compared with IAA without image classification to evaluate the influences of prior image classification in IAA. Secondly, the roles of global features (extracted from the whole image without ROIE) and local features (ROI and background features computed with ROIE) in IAA for large field images only and IAA for close-up images only are studied to clarify the role of prior ROIE in IAA.

Chapitre 3

Image aesthetic study

Sommaire

3.1 Introduction

The main goal of this chapter is to study Image Aesthetic Assessment (IAA). We are going to do a binary classification to make the distinction between high aesthetic images and low aesthetic images. Evaluating the influence of feature types (handcrafted and learned features), pre-processing operations (image classification and region segmentation) in IAA is one of the main tasks of this chapter so the binary IAA is chosen because of its simplicity.

As mentioned in the first chapter, image aesthetic is an abstract notion, it is the measure of delight or annoyance for an observer about photo fulfilling aesthetically or not the observer's expectation. The main contributions of this chapter is to answer 3 questions related to image aesthetic. The first question is "How efficient handcrafted features and learned features are in IAA ?". In order to answer this question, the performances of IAA using each feature set are estimated and compared to each other.

The second contribution is to investigate the question "Is it worthy to proceed to large field / close-up field image classification before IAA ?". The primary idea here is to assess image aesthetic of large field and close-up images separately and to consider different aesthetic features for both image categories. The LCIC methods presented in Chapter 2 are exploited

to classify large field images and close-up images. The illustration of the idea is presented in Fig. 3.1. Images are first classified as large field or close-up images. Aesthetic quality of the two categories is then assessed separately as high or low with 2 different classifiers : one designed for large field images and the other designed for close-up images. Those results are compared with the results of IAA without prior classification to evaluate the influence of LCIC in IAA.

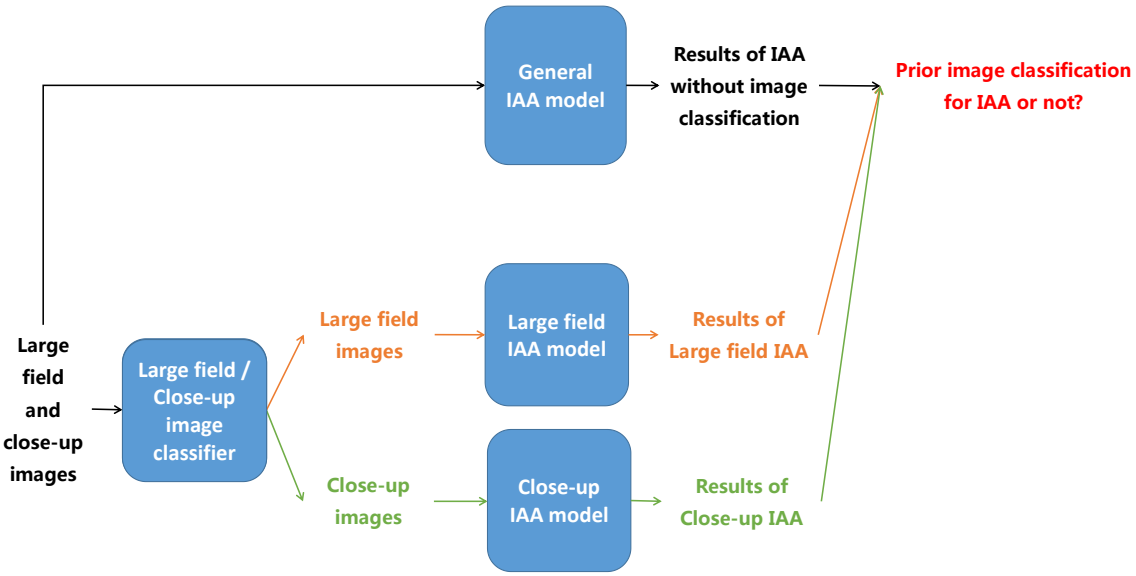


FIGURE 3.1 – The process of image aesthetic study based on LCIC results.

The last contribution is to investigate the question "Is it worthy to extract some ROIs before IAA ?". The illustration of the idea is presented in Fig. 3.2. Looking at the process, the first step is to extract the ROIs and the background from an input image. Aesthetic features are then computed from the whole image, the ROIs and the background. IAA based on each feature set (global image features, local features including ROI features and background features) are performed and compared with IAA based on both global and local features to evaluate the roles of ROIE in IAA. This problem is studied in 2 cases : IAA for large field images only and IAA for close-up images only. Large field images and close-up images are 2 typical image categories having opposite photographic rules related to ROIs and background.

Based on the evaluations of LCIC and ROIE in IAA, a new IAA model is proposed.

This chapter is organized as follows. Section 2 presents state of the art about image aesthetic. In section 3, features are defined. The study of IAA with prior image classification is described in Section 4. Section 5 presents the study of IAA with prior region segmentation. A new IAA model based on LCIC and ROIE and conclusions are drawn in Section 6.

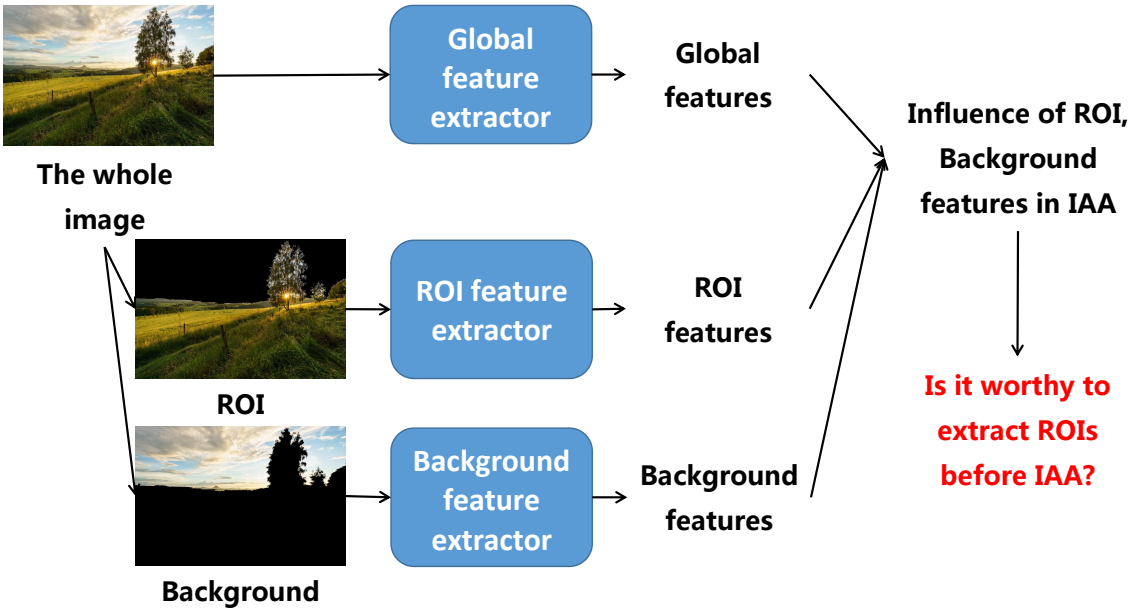


FIGURE 3.2 – The process of image aesthetic study based on ROIE results.

3.2 Image aesthetic studies : state of the art

Many attempts have been made to train computers how to automatically assess the aesthetic quality of images. Generally, there are two main phases in IAA process [DLT17]. The first one is to extract features from images : handcrafted features or learned features. In the second phase, a decision is made. The decision could be a binary classification indicating the input image as high or low aesthetic. It also could be a regression decision (returning aesthetic scores) or aesthetic ranking orders.

Following handcrafted approaches, most of studies focus on photographic rules to design aesthetic features. In [Dat+06], Datta et al. make attempts to study relations between emotions induced by pictures and low-level features of photos. Based on basic principles in photographic art, exposure of light, colorfulness, saturation, hue, the rule of thirds, familiarity measure, Wavelet-based texture, size, aspect ratio, region composition, DOF and shape convexity are studied to form 56 aesthetic features. After applying a feature selection algorithm, the 15 most relevant features are selected to train an SVM classifier performing high / low aesthetic image classification. Exploiting salient regions, Wong et al. [WL09] propose to use global features, features of salient regions and features depicting subject-background relationship to classify images as high or low aesthetic quality. In their feature set, there are 21 features designed for the whole image, 13 features designed for salient regions and 9 features representing the relations between salient regions and background. Dhar et al. [DOB11] propose to use low level features to form high level features for IAA. There are three groups of features including compositional features (presence of a salient object, rules of composition, depth of field, opposing colors), content features (presence of objects or object categories)

and Sky-Illumination features (natural illumination). An SVM classifier is trained to predict aesthetic and interestingness by using 26 high level features. Based on the color harmony of photos, Nishiyama et al. [Nis+11] propose to use bags-of-color-patterns for IAA. Local regions are first split by using grid-sampling technique. Those regions are then described by using color harmony and are quantized by using bags-of-features. The whole image is represented as a histogram of quantized features. An SVM classifier is trained based on the histogram to classify image aesthetic quality as high or low. In [Mar+11], an IAA method using a generic content-based local image signature is proposed. Bag of visual words descriptors, Fisher vector and GIST descriptors are considered to form generic content-based features. Bag of visual words descriptors, Fisher vector, gradient information are encoded by using SIFT and color information. 2 SVM classifiers are trained for binary image aesthetic classification, one with SIFT and the other with color features. The average of the 2 results is considered as the final result. Mavridaki et al. [MM15] propose to use 5 feature groups including simplicity, colorfulness, sharpness, pattern and composition to perform IAA. Their feature vector is constructed from both low and high level features computed on both the whole image and local regions. In [ASG15], Aydin et al. introduce an aesthetic signature concept and an aesthetic quality assessment method based on sharpness, depth, clarity, tone and colorfulness features. Their results prove that the aesthetic signature can help improving automatic aesthetic judgment, automated aesthetic analysis, tone mapping evaluation,...

Aesthetic is an abstract concept depending on subjective opinions and sometimes it is not easy to explain and describe it clearly so there are limitations of handcrafted features for this task. Deep learning approach might be a good solution in this case. Indeed, many researches about image aesthetic using deep learning have been introduced. Tian et al. [Tia+15] introduce a query-dependent aesthetic model with deep learning for IAA. They combine a retrieval system and a deep Convolutional Neural Network (CNN) to improve the performance of IAA. Given an input image, visual features and textual features are extracted first as the input for the retrieval system. Images in similar categories are retrieved to construct a training set for the aesthetic model. The model is then trained on the constructed training set to predict the aesthetic label for the input image. Their idea is interesting but the execution time could be an issue since whenever evaluating the aesthetic quality of an image, a retrieval task has to be executed first and the aesthetic model then has to be trained before predicting the aesthetic label. In [Lu+15], a double-column deep CNN is proposed to perform IAA. 2 parallel CNNs are used : one learning aesthetic features from the whole image and the other learning aesthetic features from local parts. Those features are then combined to classify images as high or low aesthetic quality. Additionally, style and semantic attributes are leveraged in their work. In [Kon+16], a deep CNN is proposed to rank image aesthetic based on a combination of meaningful photographic attributes (interesting content, object emphasis, good lighting, color harmony, vivid color, shallow depth of field, motion blur, rule of thirds, balancing element, repetition and symmetry) and image content. They conduct an experiment to collect aesthetic scores and photographic attributes assignments for 10,000 images. The CNN is trained to learn aesthetic features from image content and to combine them with the style attributes for rating aesthetic task. In order to evaluate aesthetic of data augmented versions, Mai et al. [MJL16] propose to use a Multi-net adaptive spatial pooling convolutional net architecture to predict aesthetic labels for images of any size. It is not similar to CNNs with regular pooling layers

where the size of the input is fixed, the input size of CNNs with adaptive spatial pooling layers is not fixed. Therefore, images of any size and ratio can be the input of those networks. Additionally, their Multi-net architecture contains many sub-networks learning aesthetic features in different image scales. Those features are then combined to decide aesthetic labels. The proposed model in that study is applied for automatic cropping by comparing aesthetic scores between the original version and the transformed versions. Focusing on the aesthetic ranking issue, Lv et al. [LT16] propose to use a pairwise-based ranking model to order photos by aesthetic quality. The main idea is to use image pairs to compute ordering information between the 2 images in each pair (in which one is more aesthetically pleasant) rather than the absolute label ("high" or "low" aesthetic quality). Image pairs are considered as the input of the model that is trained to form a ranking function. The aesthetic features in that work are extracted by an CNN with 7 layers and the output is generated by a ranking SVM. In [Wan+16a], Wang et al. introduce an CNN including 3 groups of layers to evaluate image aesthetic of multi-scenes. The first group of layers contains 4 convolutional layers pre-trained on the ImageNet dataset. The second one consists of 7 parallel groups in which each group is corresponding to a kind of scene in the CUHKPQ dataset (animal, architecture, human, landscape, night, plant and static). Each group of layers is pre-trained on the corresponding image group of the CUHKPQ dataset. The last group includes 3 fully connected layers to evaluate image aesthetic as high or low. Their model is a combination of transferred layers, scene convolutional layers and fully connected layers. Using parallel pre-trained sub-networks, a brain-inspired deep network is introduced in [Wan+16b]. The model contains 2 components : a learning attribute component and a high level synthesis component. The first one includes 17 parallel pathways in which the 3 first pathways are simply to extract hue, saturation and value information while the 14 remaining ones are convolutional sub-networks trained to determine 14 individual labels corresponding to 14 styles of the AVA dataset [MMP12]. The output of the first component is the input of the convolutional layers of the high level synthesis component. The model can be trained for binary rating prediction or ratting distribution prediction of image aesthetic.

In general, image aesthetic has been studied in various ways and prior region segmentation [WL09] ; [ASG15] ; [SS16] and prior image classification [Tia+15] have been considered. However, those studies only focus on applying prior region segmentation and prior image classification in IAA (how to exploit or apply them in IAA ? How good the performances of methods are ?) and they have not evaluated "is it worthy to perform both prior region segmentation and prior image classification for IAA ?" (How better performances are if they are applied ? How good each feature set is ? How is the role of ROIE in IAA for different image categories ? How to apply both image classification and region segmentation in IAA ?). Additionally, the question "How efficient handcrafted features and learned features are in IAA ?" still needs to be answered. In this study, we are going to tackle those problems.

3.3 Feature definition

Features in this section are defined for the purpose of evaluating the influence of prior ROIE and LCIC in IAA so 3 feature sets computed on the whole image, ROIs and background are built for IAA for all image categories (General IAA - GIAA), IAA for large field images only (Large field IAA - LIAA) and IAA for close-up images only (Close-up IAA - CIAA). Additionally, rules of photographic art are the main inspirations for designing aesthetic features either on the whole images or on local regions. However, aesthetic is an abstract concept depending on individual feelings and subjective opinions so it is not easy to describe, explain or modelize all aesthetic aspects and aesthetic characteristics. Learned features could be a good solution for this problem. Therefore, both handcrafted and deep learning based feature approaches are considered in this chapter.

3.3.1 Handcrafted features

Starting with a large handcrafted feature set built from common handcrafted features (computed from the whole image, ROIs and background based on hue, saturation, brightness, red, green and blue channels, sharpness, color saliency and contrast information) appearing in different researches [Vai+99] ; [Dat+06] ; [KTJ06] ; [LT08] ; [ASG15], the feature selection process presented in 2.3.3 is applied with 18,048 images coming from various image categories, 800 large field images and 800 close-up images to build 3 aesthetic feature sets for GIAA, LIAA and CIAA respectively. Feature vector F_h^a containing 24 features is considered for the GIAA task while 2 feature vectors : F_h^l containing 21 features and F_h^c containing 23 features are considered for LIAA and CIAA respectively. The details of the 3 feature sets are presented in Table 3.1, Table 3.2 and Table 3.3.

3.3.2 Learned feature definition

Even though the most relevant features are selected from many handcrafted features, it is possible that some aesthetic aspects have not been considered so the idea here is to use deep learning based approach to tackle the problem.

Learned features for GIAA : 3 deep CNNs are used to learn aesthetic features from the whole image, ROIs and background. A typical CNN architecture with an input layer, an output layer and 5 convolutional blocks (see the general architecture of the 3 CNNs in Fig. 3.3) is chosen. Each convolutional block has 2 convolutional layers and a pooling layer. The numbers of kernels in those blocks are 64×2, 128×2, 256×2, 512×2, 1024×2 respectively (there are 2 convolutional layers in each block). In the 4 first blocks, max pooling layers are used while a global average pooling layer is used in the last block and it is connected to a batch normalization layer before passing data to the output layer. The output layer contains 2 output neurons corresponding to the 2 classes : high aesthetic image and low aesthetic image

Features	Formula
Global features	f_1 : the mean of gradient values f_2 : the mean of brightness values f_3 : the standard deviation of brightness values f_4 : the number of main brightness bins (brightness range is split into 64 bins) f_5 : the mean of saturation values f_6 : the standard deviation of saturation values f_7 : the kurtosis of saturation values f_8 : the standard deviation of hue values f_9 : the number of main hue bins (hue range is split into 64 bins) f_{10} : the number of main colors $f_{11} = \sqrt{\sigma_{Re}^2 + \sigma_{Gr}^2 + \sigma_{Bl}^2}$ σ_{Re}, σ_{Gr} and σ_{Bl} are standard deviation of red, green and blue values f_{12}, f_{13} : the coordinate of the center point determined by gradient values f_{14}, f_{15} : the coordinate of the center point determined by saturation values f_{16}, f_{17} : the coordinate of the center point determined by brightness values
ROI and background features	f_{18} : the number of main hue bins of ROIs f_{19} : the mean of gradient values of ROIs f_{20} : the brightness contrast between ROIs and background f_{21} : the mean of gradient values of background f_{22} : the mean of brightness values of background f_{23} : the number of main saturation bins of background f_{24} : the number of main hue bins of background

TABLE 3.1 – Overview of the proposed handcrafted features F_h^a for GIAA.

Features	**Formula**
Global features	f_1 : the mean of gradient values f_2 : the standard deviation of gradient values f_3 : the mean of brightness values f_4 : the standard deviation of brightness values f_5 : the mean of saturation values f_6 : the standard deviation of saturation values f_7 : the colorfulness f_8 : the min distance to intersection points (based on the rule of thirds) determined by sharpness values f_9 : the min distance to intersection points (based on the rule of thirds) determined by color saliency values f_{10} : the min distance to intersection points (based on the rule of thirds) determined by brightness values $f_{11} = \min(f_8, f_9, f_{10})$
ROI and background features	f_{12} : the mean of gradient values of ROIs f_{13} : the mean of color saliency values of ROIs f_{14} : the mean of saturation values of ROIs f_{15} : the mean of brightness values of ROIs f_{16} : the colorfulness of ROIs f_{17} : the sharpness contrast between ROIs and background f_{18} : the color contrast between ROIs and background f_{19} : the brightness contrast between ROIs and background f_{20} : the saturation contrast between ROIs and background $f_{21} = \max(f_{18}, f_{19}, f_{20})$

TABLE 3.2 – Overview of the proposed handcrafted features F_h^l for LIAA.

Features	**Formula**
Global features	f_1 : the colorfulness f_2 : the min distance to intersection points (based on the rule of thirds) determined by sharpness values f_3 : the min distance to intersection points (based on the rule of thirds) determined by color saliency values f_4 : the min distance to intersection points (based on the rule of thirds) determined by brightness values $f_5 = \min(f_2, f_3, f_4)$ f_6 : the distribution of sharpness values f_7 : the distribution of color saliency values
ROI and background features	f_8 : the mean of gradient values of ROIs f_9 : the standard deviation of gradient values of ROIs f_{10} : the mean of color saliency values of ROIs f_{11} : the standard deviation of color saliency values of ROIs f_{12} : the mean of saturation values of ROIs f_{13} : the standard deviation of saturation values of ROIs f_{14} : the mean of brightness values of ROIs f_{15} : the standard deviation of brightness values of ROIs f_{16} : the colorfulness of ROIs f_{17} : the mean of gradient values of background f_{18} : the colorfulness of background f_{19} : the sharpness contrast between ROIs and background f_{20} : the color contrast between ROIs and background f_{21} : the brightness contrast between ROIs and background f_{22} : the saturation contrast between ROIs and background $f_{23} = \max(f_{21}, f_{22}, f_{23})$

TABLE 3.3 – Overview of the proposed handcrafted features F_h^c for CIAA.

while the input layer receives color images of size 448 × 448 (448 × 448 × 3). From an input image, 2 transformed versions are generated. In the first one, values of all pixels belonging to the background are set to 0 while all values of pixels in the ROIs are kept the same as the corresponding pixels in the input image (see Fig. 3.4(c), this is for ROI feature learning). In contrast, all pixel values of the ROIs in the second version are set to 0 while all background pixel values are kept the same as the corresponding pixels of the input image (see Fig. 3.4(d), this is for the background feature learning). The first CNN considers the original image as the input of the model to learn aesthetic features from the whole image while the second and the third models consider the first and the second transformed versions as the input to learn aesthetic features from ROIs and background respectively (see Fig. 3.3(a)).

Those deep CNNs are trained on 9024 high aesthetic images and 17,666 low aesthetic images coming from the CUHKPQ dataset [TLW13]. Those models require a very big number of samples so a data augmentation method is applied. Similarly to the data augmentation in 2.2.4.1, from the original version of a low aesthetic image, 100 transformed versions of size 448 × 448 (the resolution is good enough to keep the same label as the original version) are generated by re-scaling, padding, cropping and shifting while 200 transformed versions of size 448 × 448 are generated from the original version of a high aesthetic image by re-scaling, padding, cropping, shifting and flipping (flipped versions are added to balance the number of images in the 2 classes). Thus, the numbers of high and low aesthetic image in the training set are 1,804,800 ($9024 \times 2 \times 100$) and 1,766,600 ($17,666 \times 100$) respectively (the labels of transformed versions are set the same as the label of the original version). If the last layer of each model is removed, the 3 models become 3 feature extractors computing 1024 aesthetic features learned from the whole image F_l^g, 1024 aesthetic features learned from ROIs F_l^r and 1024 aesthetic features learned from background F_l^b respectively.

In order to compare with the handcrafted feature set F_h^a, the 24 (the same number as the number of handcrafted features for GIAA) most relevant features (F_l^a) are selected for GIAA based on feature relevance computed by the Relief method.

Learned features for LIAA and CIAA : In general, learning features directly from images often requires many samples. Although there are some datasets with aesthetic labels for all kinds of images, an aesthetic dataset for only large field images and close-up images is not available so we do not have enough data to learn aesthetic features directly. Transfer learning could be a good choice in this case. Starting with the aesthetic features $F_l^{a^*} = F_l^g \cup F_l^r \cup F_l^b$ learned in the previous part, there are 3072 aesthetic features including 1024 global features (F_l^g : features learned from the whole image), 1024 ROI features (F_l^r : features learned from the ROIs) and 1024 background features (F_l^b : features learned from the background). Those features are learned to perform GIAA for all kinds of images and we want to transfer them to focus on large field images only and close-up images only. The main idea in this case is presented in Fig. 3.5, the deep models without the last layer are considered as feature extractors to compute global features, ROI features and background features. Those computed features of large field images and close-up images only are considered as input to train new IAA models for large field images and close-up images respectively. There is a feature selection step in the

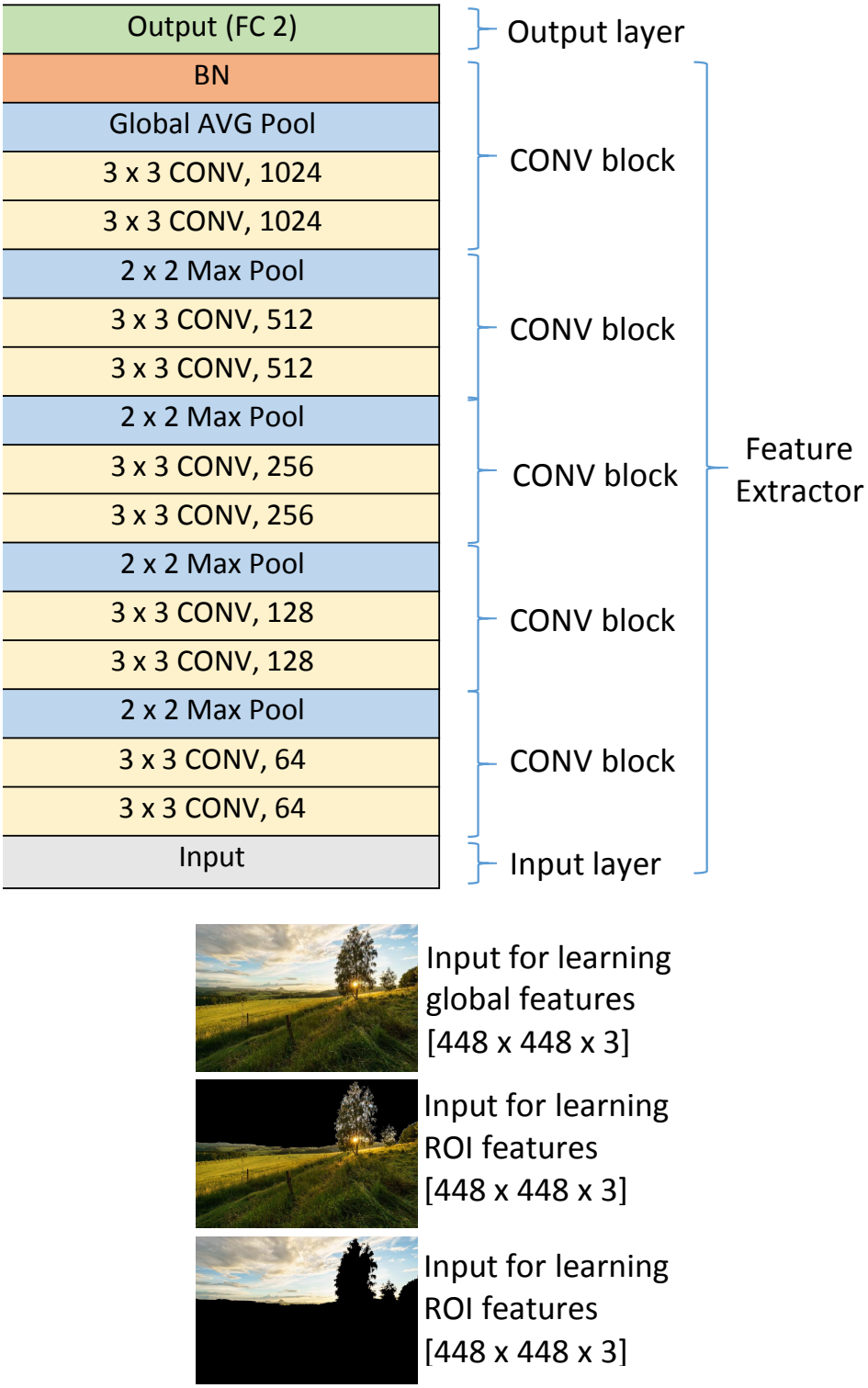


FIGURE 3.3 – The general structure of the models learning aesthetic features from the whole image, ROIs and background.

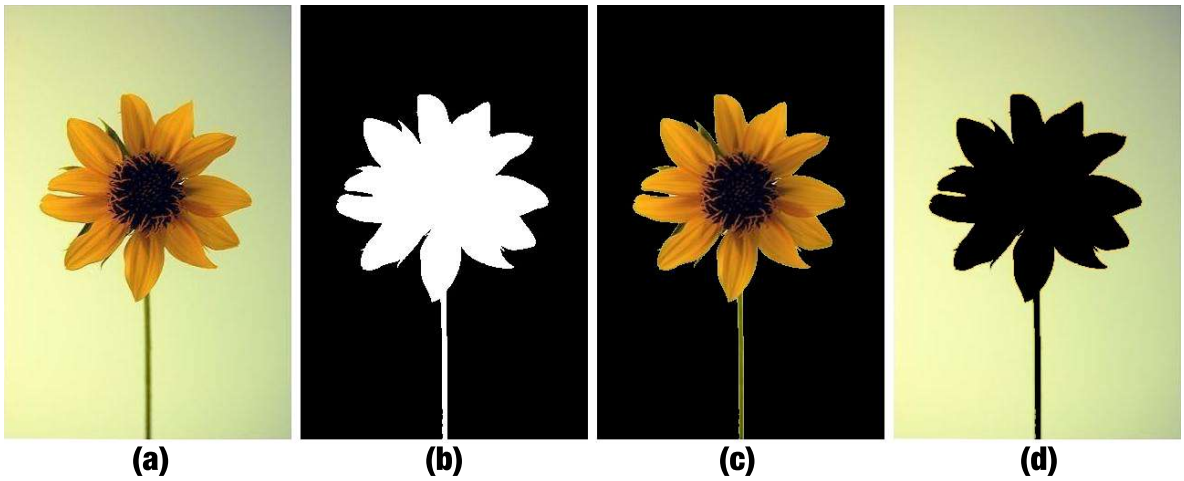


FIGURE 3.4 – Examples of the two generated versions based on ROIE. (a) The original image for global feature learning. (b) The ROI map. (c) The first version for ROI feature learning. (d) The second version for background feature learning.

process because there are 3072 learned features while the number of large field and close-up images used in this work is 2400 (1200 large field images and 1200 close-up images). It seems that the higher number of features could lead to an overfitting so it is necessary to reduce the number of learned features. The feature relevance computed by using the Relief method presented in 2.3.3 is applied to select the most relevant features in order to form the aesthetic feature vectors for LIAA and CIAA tasks. The 21 most relevant features (F_l^l) are selected from the 3072 learned aesthetic features to perform the LIAA task (the same number as the number of handcrafted features for LIAA) and the 23 most relevant features (F_l^c) are selected for the CIAA task (the same number as the number of handcrafted features for CIAA).

3.4 IAA : prior image classification or not prior image classification ?

The main question of this section are "Is it worthy to proceed to large field / close-up field image classification before IAA ?". In order to answer the question, the IAA based on the results of the prior Large field / Close-up Image Classification (LCIC) is compared with the IAA without prior LCIC. In this section, we use 2 approaches : handcrafted features and learned features to answer also the question "How efficient handcrafted features and learned features are in IAA ?".

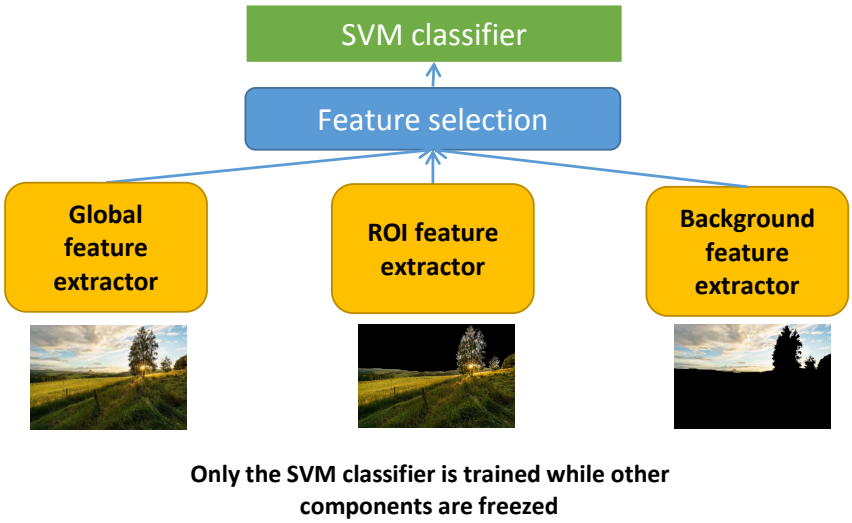


FIGURE 3.5 – Transfer learning process for Large field Image Aesthetic Assessment (LIAA) and for Close-up Image Aesthetic Assessment (CIAA).

3.4.1 Experiment and results

3.4.1.1 Dataset and setup

A part of the CUHKPQ dataset is extracted to form an aesthetic dataset with large field and close-up images only. The CUHKPQ dataset is collected mainly from DPChallenge.com website and from some other sources. All the images are labelled as high or low aesthetic. A photo is indicated as high / low aesthetic if there are at least eight of the ten viewers having the same opinion about the image aesthetic [TLW13]. There are 7 categories of the CUHKPQ dataset including animal, plant, static, architecture, landscape, human and night. Large field images are selected from the architecture and landscape categories while close-up images are extracted from the animal, plant, static and human categories (see examples in Fig. 3.6). The extracted part contains 1200 large field images and 1200 close-up images in which 50% of the images in each category are labelled as high aesthetic and the others are labelled as low aesthetic. Experiments in this section are organized on the extracted dataset. 800 large field images and 800 close-up images are selected for training and the remains (400 large field images and 400 close-up images) are used for testing.

There are main 2 experiments in this section. The first one is to perform IAA without prior image classification on both large field and close-up images using the feature vectors F_h^a and F_l^a (for all kinds of images). The second experiment is to perform the IAA with prior LCIC using the feature vectors F_h^l, F_l^l (features for large field images only) for LIAA and using feature vectors F_h^c, F_l^c (features for close-up images only) for CIAA. Those experiments are performed to answer 2 questions : "Is it worthy to perform prior image classification for IAA ?" and "How efficient handcrafted features and learned features are in IAA ?".

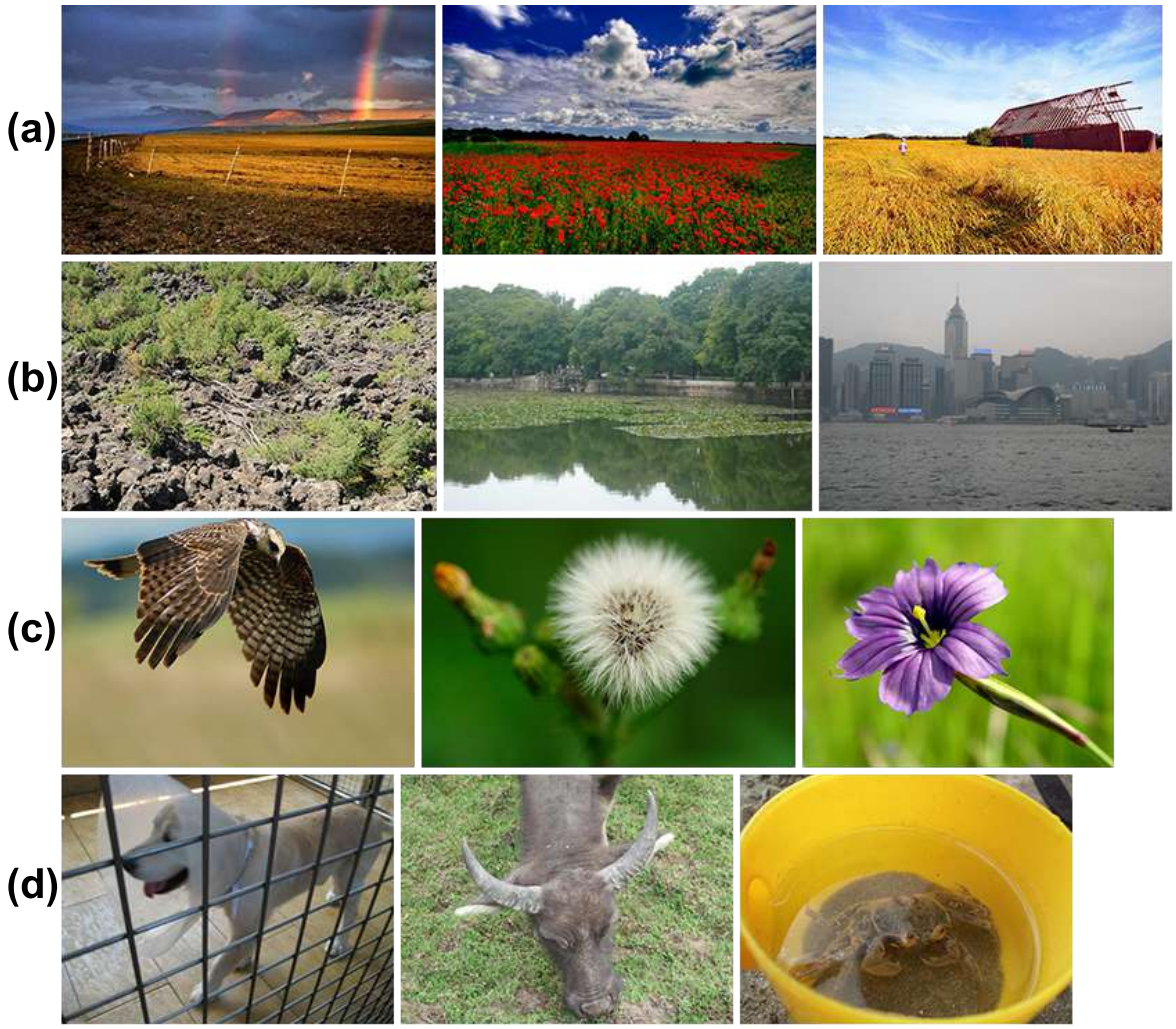


FIGURE 3.6 – Examples of high and low aesthetic images : (a) high aesthetic large field images, (b) low aesthetic large field images, (c) high aesthetic close-up images, (d) low aesthetic close-up images.

Evaluation criteria	Formula
Accuracy	$A = \frac{TP+TN}{TP+FP+TN+FN}$
Confidence interval	$I_a = z \times \sqrt{\frac{(1-A)\times A}{N}}$
Lower accuracy	$A_l = A - I_a$
Upper accuracy	$A_u = A + I_a$

TABLE 3.4 – Overview of evaluation criteria for IAA. $z = 1.96$ for 95% confidence interval and the number of samples N is 800, 400 and 400 for GIAA, LIAA and CIAA respectively. TP, FP, TN, FN are a number of images.

An SVM classifier is trained based on those feature vectors to indicate an image as high or low aesthetic. The parameters for the SVM are set as $C = 0.5$, $\gamma = auto$. Different kernels including Poly, Linear, RBF and Sigmoid are tested and only the best results with an RBF kernel are presented.

The evaluation criteria of the experiments are presented in Table 3.4. Accuracy (A), a popular evaluation criterion for classification tasks is the main criterion for the evaluation while confidence interval (I_a), the lower bound of the accuracy (A_l) and the upper bound of the accuracy (A_u) reflect the range of the accuracy. The experiments have been conducted on a PC equipped with an Intel(R) Xeon(R) W-2104 CPU 3.20 GHz, 31.7 GB memory and GPU NVIDIA Quadro P400.

3.4.1.2 Results and discussion

The results of IAA with and without image classification are presented in Table 3.5. Either with handcrafted features or learned features, the performances of IAA with prior image classification are better than the results of IAA without prior image classification (0.940 ± 0.023, 0.925 ± 0.026 for LIAA, CIAA versus 0.921 ± 0.018 for GIAA with learned features and 0.913 ± 0.028, 0.843 ± 0.036 for LIAA, CIAA versus 0.785 ± 0.028 for GIAA with handcrafted features). It appears that performing LIAA and CIAA separately using different aesthetic features could enhance the IAA performance. It could be explained that large field images and close-up images are 2 image categories having opposite photographic rules such as the composition, depth of field, focus,...so the criteria for LIAA and CIAA are not the same. Considering the relation between the 2 feature sets F_l^l (features for LIAA) and F_l^c (features for CIAA), they are almost different since there are only 3 overlapping features between the 2 feature sets. Thus, the aesthetic quality of the 2 image categories should be assessed separately using different criteria. The results and the explanation lead to a conclusion "It is worthy to proceed to large field / close-up field image classification before IAA".

Moving to the second question "How efficient handcrafted features and learned features are in IAA ?", in the general case (GIAA for all image categories), learned features are better than handcrafted features since the GIAA results with learned features and handcrafted features are 0.921 ± 0.018 and 0.785 ± 0.028 respectively. As mentioned in the previous part, image

Feature vector	A	I_a	A_l	A_u
GIAA - IAA without image classification				
F_h^a	0.785	0.028	0.757	0.813
F_l^a	0.921	0.018	0.903	0.939
LIAA - IAA for large field images only				
F_h^l	0.913	0.028	0.885	0.941
F_l^l	0.940	0.023	0.917	0.963
CIAA - IAA for close-up images only				
F_h^c	0.843	0.036	0.807	0.879
F_l^c	0.925	0.026	0.899	0.951

TABLE 3.5 – Evaluations of IAA with and without image classification using handcrafted and learned features.

aesthetic is an abstract concept depending on human perception and individual feeling so understanding and defining all aesthetic aspects are not easy. However, handcrafted aesthetic features are designed to reflect aware aesthetic aspects so it is impossible to design handcrafted features representing inexplicable aesthetic aspects. On the contrary, deep models can learn complex and inexplicable aesthetic features so we can find some similarities between image aesthetic notion and learned features. It could be the reason why the results with learned features are better than the ones with handcrafted features. Considering the second case of IAA for a particular image category (large field images only or close-up images only), the results of LIAA and CIAA with learned features and handcrafted features are 0.940 ± 0.023 versus 0.913 ± 0.028 and 0.925 ± 0.026 versus 0.843 ± 0.036 respectively. According to those results, the final conclusion is archived : learned features are very efficient and they are better than handcrafted features for IAA. The following section focuses on learned features only because of their higher performances.

The summary of the experiment results is presented in Fig. 3.7.

3.5 IAA : with or without prior region segmentation ?

The main goal of this section is to evaluate the role of ROIE in IAA. The role of ROIs is not always the same for each image so the influence of ROIE in IAA for a particular image category (large field images only or close-up images only) is going to be considered. The 2 learned feature sets F_l^l (for LIAA) and F_l^c (for CIAA) presented in the previous section are analyzed to estimate the influence of ROIE in IAA.

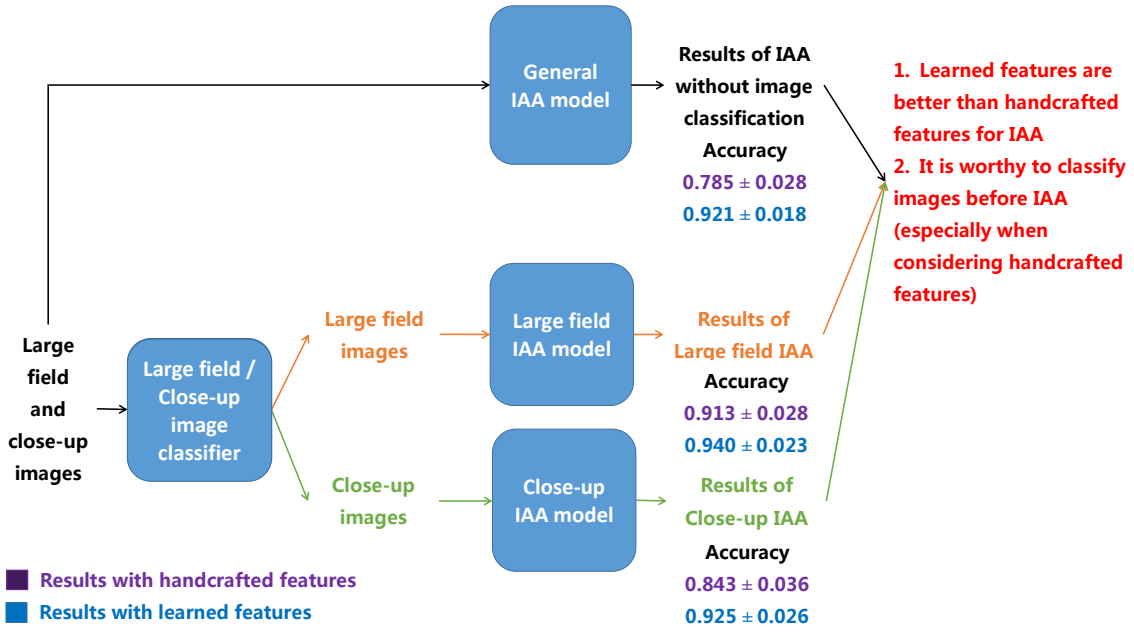


FIGURE 3.7 – Summary of the experimental results of IAAs with and without prior large field / close-up image classification.

3.5.1 Experiment and results

There are 2 main tasks in this part. Firstly, the distribution of ROI and background features (RB features) in each feature set (F_l^l and F_l^c) is analyzed to have an overall view about the role of ROIE in LIAA and CIAA. Secondly, IAA using RB features is compared with IAA using global features and with IAA using both global and RB features to estimate how ROIE affects IAA.

3.5.2 Dataset and setup

The experiments of LIAA and CIAA using the feature sets F_l^l and F_l^c respectively are performed on 1200 large field images and 1200 close-up images (the same as the dataset of the experiments of LIAA and CIAA in the previous section) in which 800 large field images and 800 close-up images (50% of the images in each category are labelled as high aesthetic and the others are labelled as low aesthetic) are used for training while the remains are used for testing.

As done before, the parameters of the classifiers are set as $C = 0.5$, $\gamma = auto$ and different kernels are tested and only the best results are presented. The main evaluation criterion is the accuracy. The range of the accuracy is presented by the confidence interval, the lower bound of the accuracy and the upper bound of the accuracy.

Feature set	The number of	
	Global features	RB features
F_l^l	21	0
F_l^c	18	5

TABLE 3.6 – The number of global features, RB features (ROI features and background features) in the 2 feature sets F_l^l and F_l^c for LIAA and CIAA respectively.

Feature vector	A	I_a	A_l	A_u
F_l^c	0.925	0.026	0.899	0.951
F_g^c	0.908	0.028	0.880	0.936
F_{rb}^c	0.868	0.033	0.835	0.901

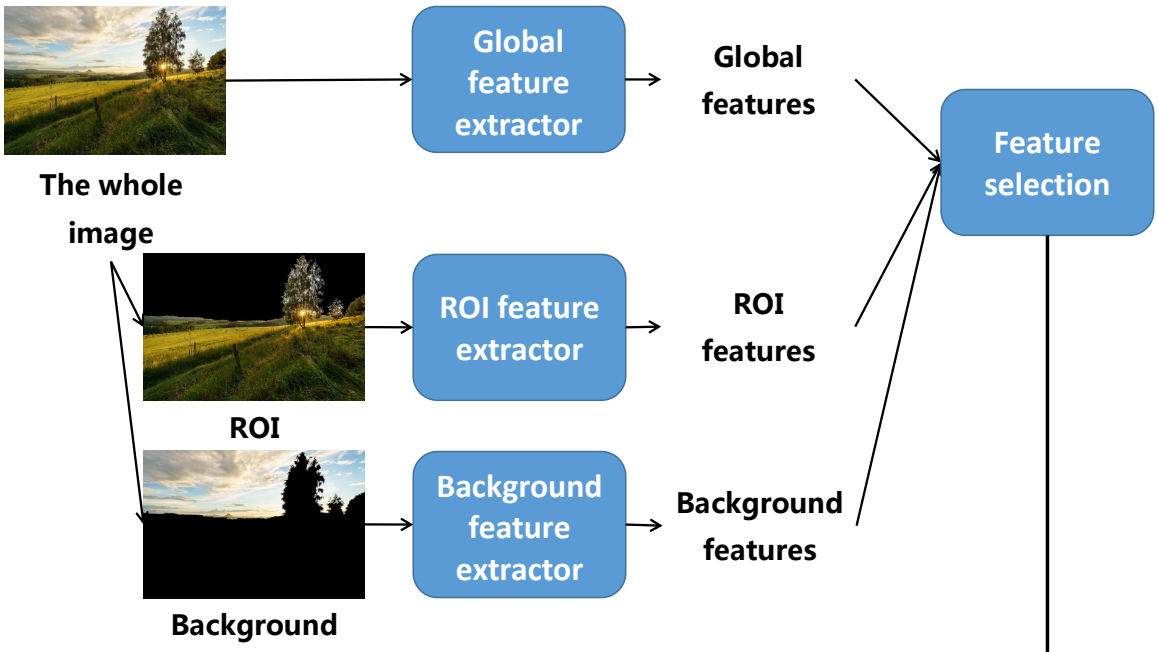
TABLE 3.7 – Evaluation of CIAA using global features, RB features (ROI and background features) and both global features and RB features.

3.5.2.1 Results and discussion

Firstly, Table 3.6 shows the number of global features and RB features (ROI features and background features) in each feature set (F_l^l and F_l^c). It appears that the role of ROIE in IAA is not the same for all image categories. In the case of close-up images, ROIE has the most significant role in IAA since the number of RB features in F_l^c is the highest (5 features). In contrast, there is no RB feature in the feature set F_l^l for LIAA. The reason probably is that the content of a large field photo is a large scene (as the name of the category) so viewers often pay attention to the whole large scene including both ROIs and background. Therefore, the influence of ROIE in LIAA is not significant so LIAA is skipped in the next analysis.

Secondly, the evaluations of global features (F_g^c : global features in F_l^c) and RB features (F_{rb}^c : ROI and background features in F_l^c) for CIAA are presented in Table 3.7. The results are quite interesting since with only 5 RB features, the obtained classification accuracy is very impressive (0.868 ± 0.033). The combination of 5 RB features and 18 global features helps increasing the IAA performance from 0.908 ± 0.028 to 0.925 ± 0.026. The background of close-up images is often blur to highlight the main close-up object regions (sharp regions with high contrasted colors - ROIs) so viewers often pay more attention on ROIs. It explains why ROIs have significant influence on aesthetic quality of close-up images. According to those results, it appears that it is worthy to extract ROIs before assessing aesthetic quality of close-up images.

The summary of the experiment results is presented in Fig. 3.8. In general, the role of ROIE in IAA is various since the influence of ROIE in IAA for large field images is insignificant while ROIE helps improving the IAA for close-up images. The answer to the question "IAA : prior region segmentation or not ?" might depend on the considered situation.



Features	Accuracy
Features for LIAA (IAA for large field images only)	
21 global features	
No RB feature (ROI and background features)	
Features for CIAA (IAA for close-up images only)	
18 global features	0.908 ± 0.028
5 RB feature	0.868 ± 0.033
(23 global + 1 RB) features	0.925 ± 0.026

"ROIE before IAA or not" depends on the considered situation

FIGURE 3.8 – Study summary about IAA with prior ROIE.

3.6 Conclusion

In this chapter, the main works were to study IAA with and without prior image classification or region segmentation. Firstly, the experimental results prove that classifying images before performing the IAA can enhance the IAA performance. Secondly, performing prior ROIE before IAA or not depends on the image type. Based on the obtained results, we propose an IAA model based on LCIC and ROIE. Fig. 3.9 presents the idea of the proposed model. Images are first classified as large field images and close-up images. Then, large field images are assessed as high or low aesthetic quality by a classifier based on global features only. On the contrary, ROIs and background are extracted from close-up images to compute ROI features and background features. Those features are then combined with global features to make the distinction between high and low aesthetic close-up images. Fig. 3.9 also shows the performances of the model compared with IAA without image classification and region segmentation. Firstly, it appears that image classification helps improving the IAA performances by assessing aesthetic quality of large field images and close-up images separately. Secondly region segmentation helps for CIAA especially in the case of handcrafted features. Both handcrafted features and learned features have been considered in this chapter and unsurprisingly learned features are more efficient.

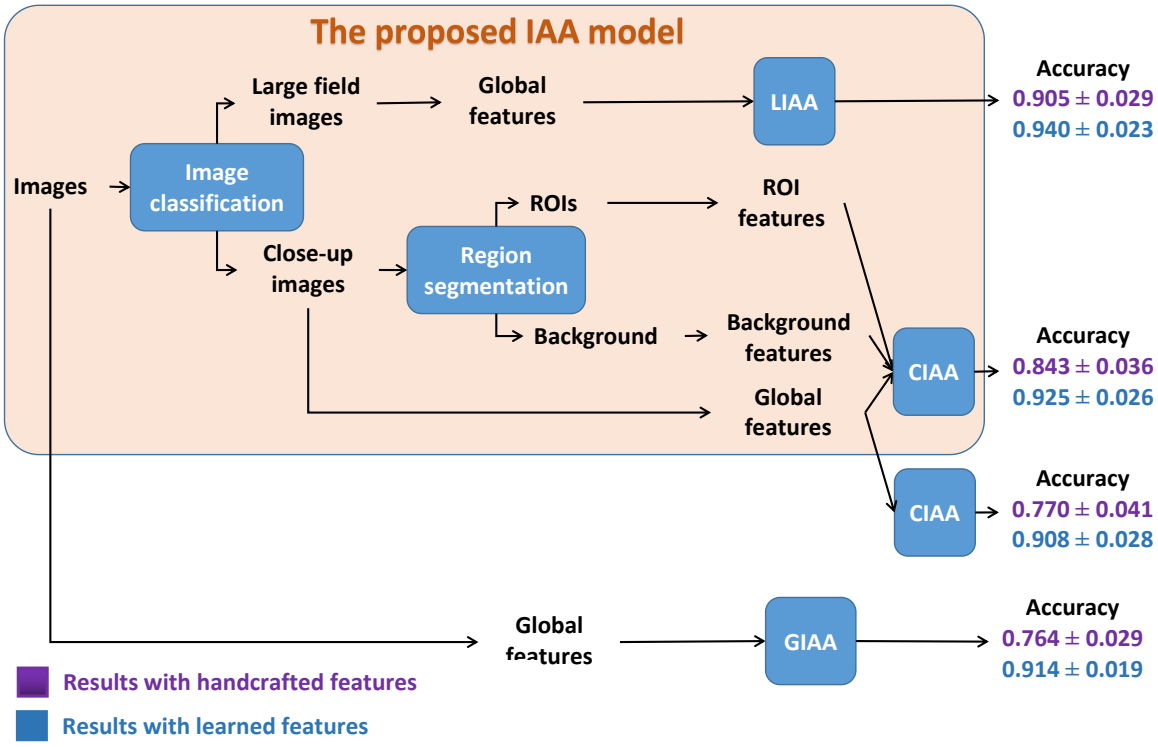


FIGURE 3.9 – Proposed algorithm for IAA.

Chapitre 4

Image naturalness study

Sommaire

In this chapter, we are going to address the second aspect of image quality : "image naturalness". Nowadays, images can be obtained in various ways such as capturing photos in single-exposure mode, applying Multiple Exposure Fusion (MEF) algorithms to generate an image from multiple shoots of the same scene, mapping High Dynamic Range (HDR) images to Standard Dynamic Range (SDR) images, converting raw formats to displayable formats, or applying post-processing techniques to enhance image quality, aesthetic quality,... When looking at those processed photos, one might have a feeling of unnaturalness, the feeling that something is wrong in the photo (see examples in Fig. 4.1).

This chapter deals first with the problem of developing a model to estimate if an image looks natural or not to humans and the second purpose is to try to understand how the unnaturalness feeling is induced by a photo : Are there specific unnaturalness clues in images or is unnaturalness a global feeling about the whole photo ? The study focuses on SDR images, especially on Tone Mapped Images (TMIs). The first contribution of the chapter is the setting of an experiment gathering human naturalness opinions about 1900 SDR images mainly obtained from tone mapping operators. Based on the collected data, the second contribution is to study the efficiency of different feature types including handcrafted features and learned

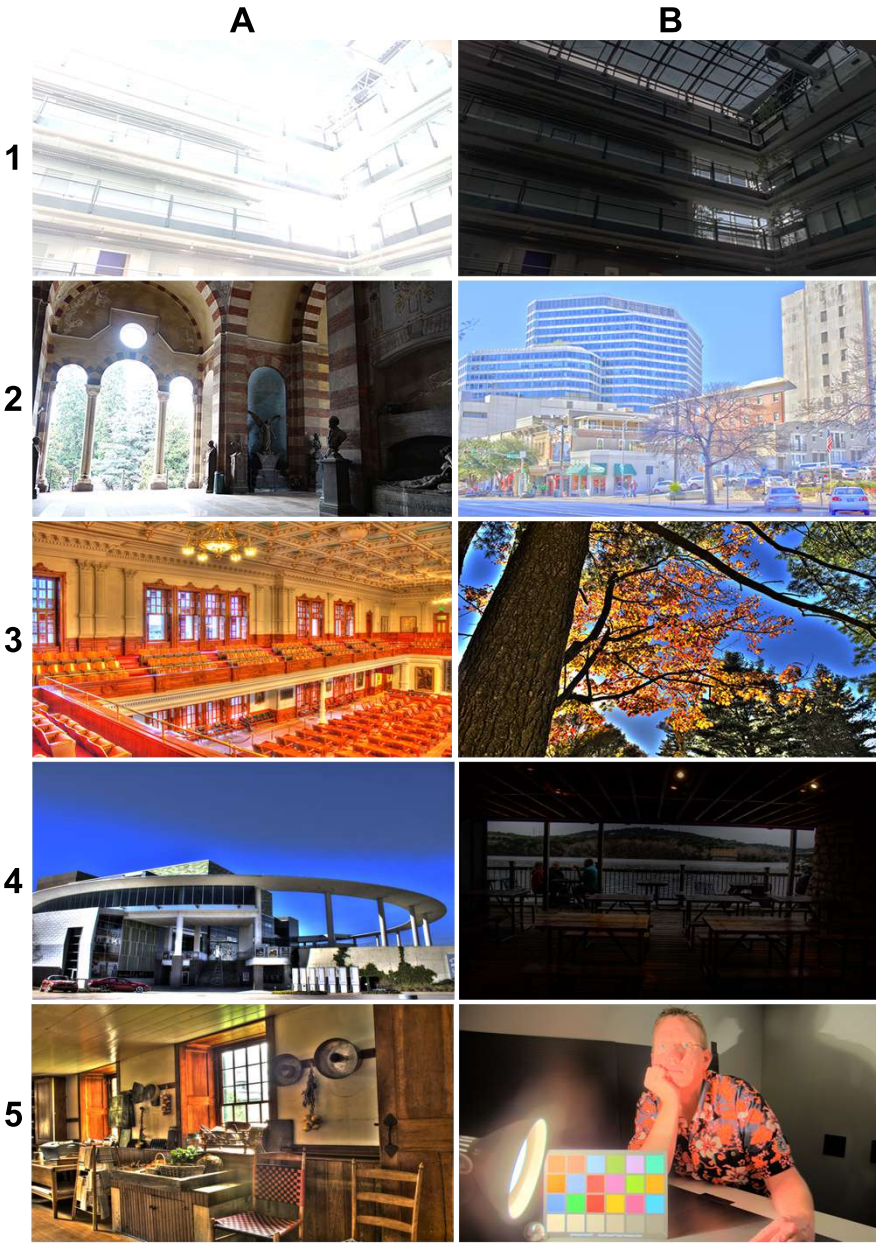


FIGURE 4.1 – Examples of artifacts. 1A : Over exposure, lost details. 1B : Under exposure, lost details. 2A : Too high contrast. 2B : Too low contrast, incorrect color reproduction. 3A : Bloom effect, incorrect color reproduction. 3B : Hallow effect, incorrect color reproduction. 4A : over saturation. 4B : under exposure. 5A : over saturation. 5B : incorrect color reproduction, halo shape of light.

features for image naturalness analysis. A binary classification model is then developed based on the determined features to classify if an image looks natural or unnatural.

4.1 Introduction

In recent years, more and more new camera models, photography techniques and image processing applications have been introduced to consumers. Three emphasises should be mentioned : High Dynamic Range (HDR) images, Multiple Exposure Fusion (MEF) algorithms and Tone Mapping Operators (TMOs). The dynamic range of images is the ratio between the highest and lowest luminance values. The dynamic range of irradiance in real scenes possibly reaches 100,000,000 :1. The human eye can perceive the dynamic ranges from 10,000 :1 to 1000,000 :1 (depending on circumstances) while a normal display is able to present a Low Dynamic Range (LDR, LDR and SDR are considered as the same concept in recent years) from 100 :1 to 300 :1 [See+04] ; [Rei+10] ; [KBK11]. As a consequence, the luminance range of scenes displayed on standard screens is narrower than that of real scenes and it is also lower than the dynamic range perception of human eyes. In the past, the problem of high dynamic range was caused by the camera sensors and the display devices. The camera sensors were not able to cover the whole irradiance range of real scenes. Nowadays, the capability of professional camera sensors has increased and those sensors can capture HDR of almost normal scenes (14 stops of dynamic range - the dynamic range is 2^{14} :1). And when the dynamic range is too high to be covered (for example, 20 stops of dynamic range) or with a non-professional camera, Multiple Exposure Fusion (MEF) algorithms can be used to help covering the whole range. MEF is a technique generating an image from multiple shoots taken under different exposures for a given scene [DM97] ; [Gos05] ; [MKVR09] by using fusion algorithms (see examples in Fig. 4.2). The MEF technique helps an image having a higher dynamic range than that of an image taken with a fixed exposure.

On the side of display devices, the work is in progress. Some of new commercial devices are able to present irradiance peaks around 1,000 $cd\ /\ m^2$ and black levels less than 0.05 $cd\ /\ m^2$ (the dynamic range is 20,000 :1). Especially, some special models used in research can reach the highest luminance value of 10,000 $cd\ /\ m^2$. Although the dynamic range of new display devices is quite high, it is still quite modest when compared to the dynamic range of real scenes and the perception range of human eyes [Bas+19]. Thus, nowadays the problem of high dynamic range images is mainly related to display devices.

8 bit data is currently used to present images displayed on standard screens. Although there is no direct relation between bit-depth and dynamic range, it is necessary to use more steps (more bits) to present a higher dynamic range. A pixel of any HDR image is represented by 3 colors and each color is coded by 10 bits, 12 bits, 16 bits or 32 bits. Although some new monitor models (HDR monitors) are able to display a high dynamic range content (20,000 :1), most of the popular display devices are SDR screens that are able to display only SDRs of irradiance. Thus, it is necessary to map HDR images to SDR format before display on SDR screens. To perform this task, many Tone Mapping Operators (TMOs) have been proposed [Kha+18] ;

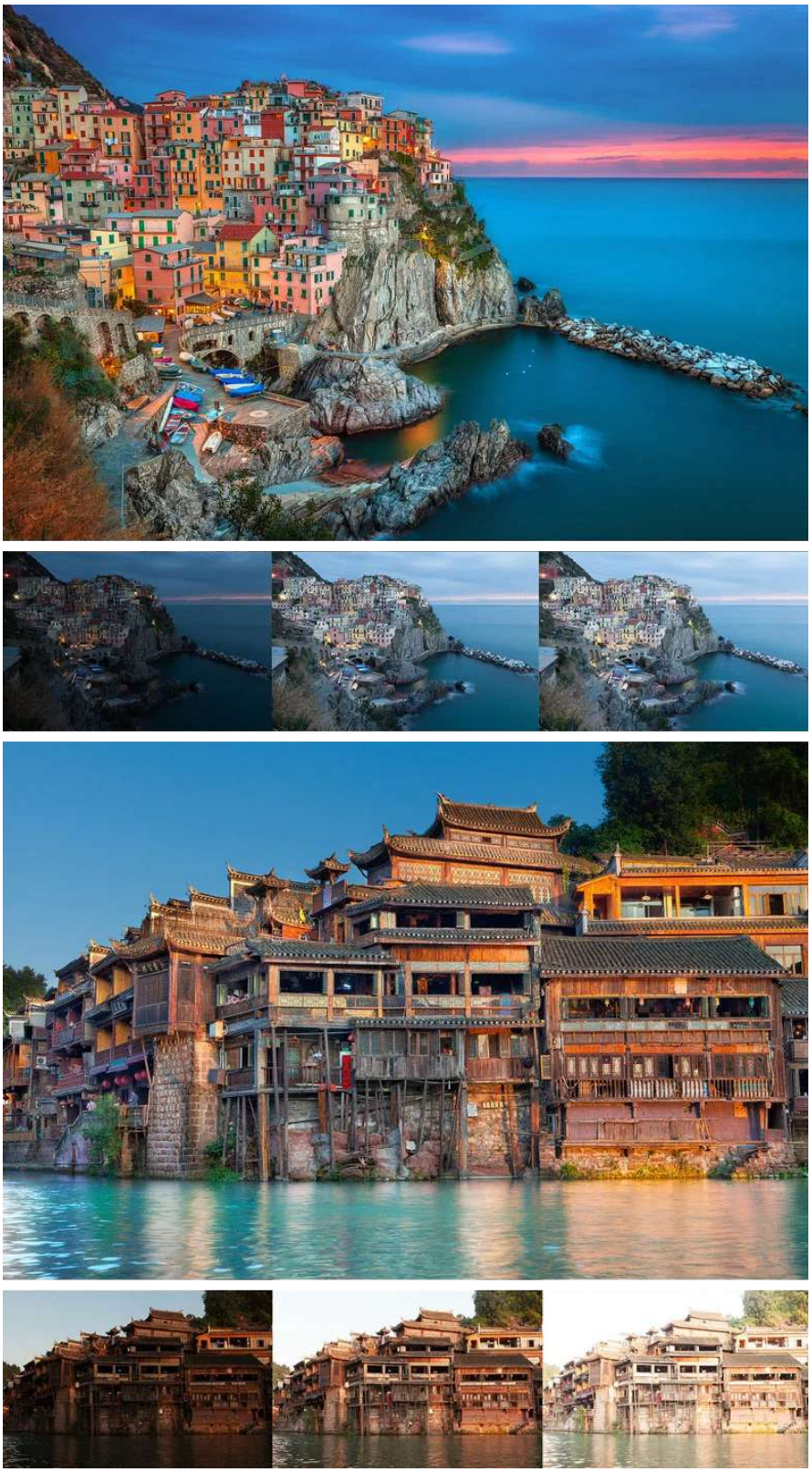


FIGURE 4.2 – Examples of MEF. The first and the third rows present images generated with the multi-exposure images of the second and the fourth rows respectively. (image source : https ://petapixel.com).

FIGURE 4.3 – Examples of post processing methods. The first column contains original images (produced directly by cameras) while the second one presents the corresponding post-processed images. (image source : https ://petapixel.com).

[KBK11] ; [Rei+02] ; [Ash02] ; [FLW02] ; [DD02] ; [LRP97]. Generally, TMOs map colors of HDR images from a HDR (from 10,000 :1 to 1000,000 :1) to a LDR (from 100 :1 to 300 :1), this process can be considered as a range compression process. Beside this, in order to correct the colors of images or to create special effects, some post-processing algorithms can be applied on SDR images. For example, in the first row of Fig. 4.3, post-production colors and contrast enhancements have been used to produce the image on the right. In the post-processed image of the second row, orange sky and sun-rays have been created by Adobe CameraRaw and Photoshop respectively. Additionally, an exposure enhancement algorithm has been used in that photo. In the third row, Nik Color Efex Pro and Photoshop have been used to enhance the colors and to create dodge and burning effects. In the fourth row, the right image is obtained by applying a multiple exposure blending algorithm. In the last case, VSCO and Nik Color Efex Pro are used to enhance colors, the contrast and the color temperature are also corrected to obtain the right photo.

One problem of SDR images obtained by using those algorithms might be the loss of naturalness or the appearance of unnaturalness (see examples in Fig. 4.1).

In this research, the naturalness concept is focused. On the one side, an image is considered as natural if the appearance of the image looks familiar to a human observer (it makes the observer have the feeling that the photo is a faithful representation of the scene). On the other side, if the observer has the feeling that something in the photo is wrong (due to color appearance, abnormal details or more subtle changes) so that the appearance of the photo does not look faithful, the photo is considered as unnatural. In this work, the naturalness concept is not supposed to be related to the image content itself. For example, augmented images are considered as natural in this study (see Fig. 4.8). The research focuses on collecting naturalness opinions from viewers to design features representing naturalness and unnaturalness. This work is neither about image aesthetic assessment [DLT17] nor image quality assessment [MEMS14].

In this study, there are two main contributions. The first one is an experiment of subjective Image Naturalness Assessment (INA) without references. The experiment is conducted thoroughly at the laboratory with a set of SDR images obtained in various ways. The second contribution is the study of different features including handcrafted, shallow and deep learned features (learned from shallow and deep CNNs respectively) for the INA task with the hope to define the best features describing naturalness / unnaturalness [LE+20].

This chapter is organized as follows. Section 2 presents the state of the art about image naturalness. Section 3 introduces the experiment of subjective INA and the dataset that has been collected. In section 4, feature definition and feature selection for INA are described. Section 5 presents the results of automatic natural / unnatural SDR image classification. Section 6 generalizes the understanding about image naturalness. The relations between image naturalness and image aesthetic are clarified and discussed in section 7.

[RBF95]'s and [Rid96]'s studies
Image naturalness is defined as a high degree of correspondence to (memorized) reality. High quality images should be considered as natural.
Naturalness indexes : - Chromatic variation. - Hue variation. - Saturation variation. - Lightness variation.

[CS05]'s study
Image naturalness is defined as a degree of correspondence between a scene (seen directly) and the same scene in photos based on some criteria : brightness, contrast, colour reproduction, reproduction of details, simulation of glare, visual acuity. and artifacts.
Naturalness indexes : - Brightness. - Contrast. - Colour reproduction. - Reproduction of details. - Reproduction of shadow details. - Simulation of glare. - Visual acuity. - Artifacts.

[Cho+09]'s study
Image naturalness is defined as the degree of correspondence between a photo displayed on a device and the memories about the real-life scene.
Naturalness features : - Memory colors of skin, grass and sky. - Sharpness. - Colorfulness. - Reproduction of shadow details.

[Gu+16]'s study and [YMB17]'s study
Natural images are images that can be obtained from a camera - these include pictures of man-made objects as well as forest/natural environments and the remains are considered as unnatural.
Naturalness feature : it is calculated based on standard deviation and mean of pixel values and a statistic of natural images.

[Jia+18]'s study
Image naturalness definition is based on exposure of images. Over or under exposure images are considered as unnatural images while normal exposure images are considered as natural.
Naturalness features are calculated based on luminance and yellow intensities.

TABLE 4.1 – Overview of naturalness definitions, indexes and features in previous studies.

4.2 Image naturalness studies : state of the art

In the literature, different definitions of image naturalness have been given. In [Cho+09], image naturalness is defined as the degree of correspondence between a photo displayed on a device and the memories about a real-life scene. An experiment is conducted with 13 observers, 8 color images and 22 manipulations of them to gather the perceived naturalness. The perceived naturalness is then compared with a naturalness index based on sharpness, colorfulness and reproduction of shadow details, memory colors of skin, grass and sky. In [RBF95] and [Rid96], image naturalness is defined as the same as in [Cho+09] but they propose that high quality images should be considered as natural. By analyzing the chromatic, hue, saturation and lightness variations, they point out the significant roles of those factors in image quality and image naturalness. But those studies only focus on evaluating the impacts of some factors on naturalness instead of finding factors affecting photo naturalness. In [CS05], image naturalness is defined as a degree of correspondence between a scene (seen directly) and the corresponding scenes in photos based on some criteria : brightness, contrast, colour reproduction, reproduction of details, simulation of glare, visual acuity and artifacts. An experiment is conducted to evaluate the naturalness of SDR images generated by 14 different TMOs with a human naturalness assessment experiment with references. The real scene and the tone-mapped version of an HDR image of the same scene are shown to the observers. The observers have to give a subjective score (in range $[0,\dots 10]$) for the 5 criteria including brightness, contrast, visibility, reproduction of details and reproduction of colors. Based on the subjective scores, the TMOs are compared.

Besides, some naturalness features have been proposed for tone mapped Images Quality Assessment (IQA) in few studies. In [Gu+16] ; [YMB17], naturalness is mentioned as a factor to assess image quality since it is considered as a feature in a feature set for IQA. In those researches, natural images are images that can be obtained from a camera - these include pictures of man-made objects as well as forest / natural environments while other images are considered as unnatural. It is computed based on statistics with 3000 natural images. Naturalness is considered as the fitness of the standard deviation and the mean of pixel values to a Gaussian function and a Beta probability density function. In another study, Jiang et al. [Jia+18] define naturalness features based on the differences of normal exposure images and abnormal (over or under) exposure images. Simply, over or under exposure images are considered as unnatural images in that research. Naturalness features are based on the luminance and yellow values. Those features are then used with details features and aesthetic features for tone-mapped image quality assessment. In those studies, it is concluded that naturalness plays a role in tone-mapped image quality assessment but naturalness is mentioned as a factor and there is no clear definition, conclusion or evaluation about the consistency of the naturalness. Table 4.1 presents an overview of the naturalness features used in previous studies.

The definition of naturalness in this study is related to both obvious clues such as contrast, reproduction of detail and colors, bloom, halo and dark band effects,... (see Fig. 4.4) generated by TMOs, MEFs, processing algorithms AND

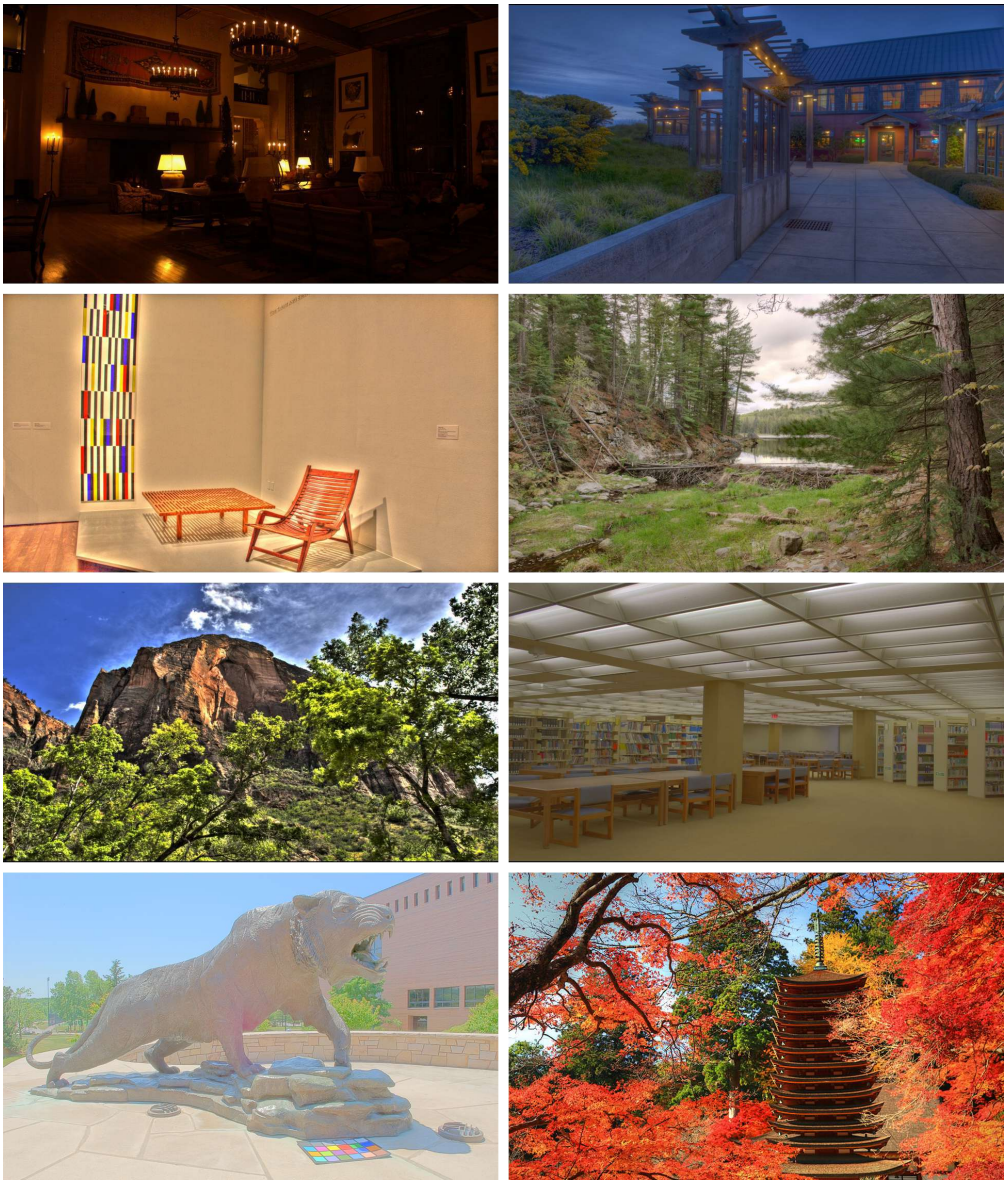


FIGURE 4.4 – Examples of unnatural images. In the left column, the images have clear artifact signs. The brightness in the first image is too low, there are halos surrounding the objects in the second image, the color saturation in the last one is too high. In contrast, when the observers look at the images of the right column, they have the feeling that the images are unnatural without being able to explain easily why.

the individual feeling of naturalness / unnaturalness related to image memory. As a consequence, unnaturalness might be caused either by obvious localized clues or by a global impression when looking at the whole image. In Fig. 4.4, images in the left column give obvious unnatural clues examples while images in the right one show examples in which unnaturalness is related to a global feeling. In the state of the art about image naturalness, it is worthy to notice that none of those studies has the same definition of naturalness and their purposes are different from the purpose of this work. Moreover most of the naturalness features mentioned in previous researches are handcrafted features. But in the naturalness concept, we think that there is also an abstract part related to individual memories which cannot be precisely described and inferred by handcrafted features. As a consequence, there is probably a need about naturalness features learning and the respective influence of handcrafted and learned features on INA is still an open question.

4.3 Experiment of subjective image naturalness assessment

There are few research about image naturalness and one important challenge is that labelled natural / unnatural datasets are not available. There is a need of collecting such data so the first step before studying image naturalness is to organize an experiment of subjective INA without references. The description of the conducted experiment includes the image sources, the experiment design, the experiment process, the observers, the experiment results and the naturalness dataset built from the data collected from the experiment.

4.3.1 Image sources

The dataset contains 2727 SDR images coming from 3 main sources. The first one is 624 SDR images mapped from 208 HDR images coming from different sources. HDR images are not easy to collect and the number of images in each dataset is often small, so 7 HDR datasets (Debevec's [DM97], Fairchild's [Fai07], Cadik's [Čad+08], Narwaria's [Nar+13], Yeganeh's [YW13], Korshunov's [Kor+14] and Krasula's [Kra+17] datasets) are used in this research. Those HDR images are mapped to SDR images by using different TMOs including Reinhard's [Rei+02] (based on global contrast), Ashikhmin's [Ash02] (using local contrast) and Khan's [Kha+18] (based on histogram and human visual system) algorithms. In order to focus on both naturalness and unnaturalness, there is a need of considering not only a well known TMO like Reinhard's TMO but also an TMO generating artifacts like Ashikhmin's TMO and an TMO generating both natural and unnatural images like Khan's TMO.

The second image source includes 1,811 SDR images of ESPL-LIVE dataset [Kun+17]. It includes 747 SDR images mapped from HDR images by using 4 TMOs [Rei+02] ; [LRP97] ; [FLW02] ; [DD02], 710 images grenerated directly from multi-exposure shoots by using 5 MEFs [Kun+17] ; [RC09] ; [PK10] ; [PSA16] and 354 images gained after applying 2 post-processing algorithms [Kun+17].

The last source contains 292 images including single-exposure, tone-mapped and post-processed images downloaded from Flickr website. The contents of the images are real world scenes including landscape, building, objects, people,... and they are taken under indoor, outdoor, day time, night time conditions.

4.3.2 Experiment setup

4.3.2.1 Experiment design

The experiment was conducted at GIPSA Lab, France where the experimental conditions are controlled according to the ITU BT-500 [1] for a subjective experiment. 2 main types of experiments are proposed in ITU BT-500 : in absolute or relative terms. The experiment in this study is an absolute binary experiment. In detail, observers are asked to decide whether an image is natural or unnatural and they perform the experiment by interacting with an interface displayed on a 24 inch (16 :10) Samsung display (see Fig. 4.6). The resolution and color profile of the display have been set to 1920 × 1200 pixels and sRGB respectively. The peak brightness of the display is 250 cd / m^2. It is connected to a computer exporting a 32 bit color signal. The display and the computer are placed in an experimental room where the light conditions are controlled thoroughly. The distance from observers to the display was fixed to 0.7 meter. Although the number of observers in the laboratory experiment is lower than that of some online crowd surveys, the thorough control of experimental conditions is the compensation ensuring the reliability of the experiment results.

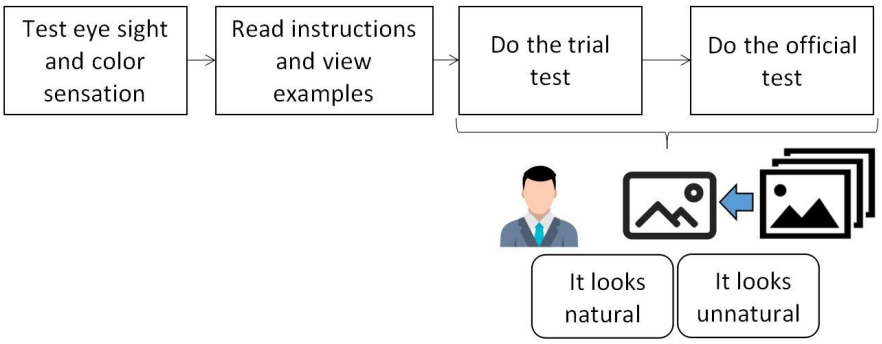


FIGURE 4.5 – The process of the experiment for an observer. There are 4 main steps including testing eyes, reading instructions, doing the trial test and doing the official test.

4.3.2.2 Experiment process

The process of the experiment for an observer is described in Fig. 4.5. Before starting the experiment, the observer performs an eye sight and a color sensation tests. The observer then

1. https ://www.itu.int/rec/R-REC-BT.500/fr

FIGURE 4.6 – The interface for assessing image naturalness. There are two options for observers : the image looks natural or the image looks unnatural. The decision is made by clicking one of the two buttons or pressing a corresponding key on the keyboard.

reads the instructions, views some examples and performs a trial experiment to understand the experiment precisely and to be familiar with the interface of the experiment. He / she is instructed to focus on image naturalness rather than image quality or image aesthetic. In the trial phase, he / she has to evaluate the naturalness of five photos covering different causes of unnaturalness (they were pre-evaluated) and not belonging to the official experiment phase. Each turn, only one photo is showed to the observer during a short time. As explained previously, naturalness is an abstract concept not so easy to define precisely to each observer since there is probably an unconscious part in the evaluation process. As a consequence, the observers have been asked to quote each image in a binary way : natural or unnatural. Therefore, there are only two choices : the photo looks natural or it looks unnatural to the observer (see Fig. 4.6). The observer can click a button on the interface or use the keyboard to enter his/her decision. Although the maximum time for evaluating an image is 7 seconds, the actual time in the experiment ranges from 3 to 5 seconds per photo. After giving the subjective evaluation, an uniform gray background is displayed for 1 second and the next image is then presented automatically to the observer. In the next step, the official experiment is performed in the same way as the trial experiment but the number of photos is higher. In the official phase, the number of evaluated photos per observer is 380. The total performing time per subject ranges from 25 to 30 minutes.

4.3.2.3 Observers

There were 45 people participating in the experiment which have quoted 1900 images among the 2727 available images. The number of men and women are 33 and 12 respectively. Among the 45 observers, 33 observers are familiar with image processing. The observers' ages range from 18 to 57. The average and the standard deviation of their ages are 26.2 and 7.53 respectively. The results show that 100 percent of them have normal or corrected to normal vision at that time.

4.3.3 Experiment results and the naturalness dataset construction

17,100 no reference subjective evaluations of photo naturalness were collected from 45 observers for 1900 SDR images. Each SDR image has been assessed by 9 observers. The distributions of the evaluations with respect to each transformation method are presented in Table 4.2. The distributions of the different groups are various. Some transformation methods receive a significant difference between the number of positive evaluations (assessing an image as natural) and the number of negative evaluations (assessing an image as unnatural) such as Durand's method (179 versus 1063), Surreal effect (117 versus 1062), Grunge effect (47 versus 835), Ashikhmin's TMO (230 versus 1372). In contrast, the difference in the numbers of positive evaluations and the number of negative evaluations is in-significant for Pece's method (422 versus 377), Khan's method (732 versus 870). In other cases, there is a slight difference between the number of positive evaluations and the number of negative evaluations : Reinhard's method (1549 against 725), Fattal's method (260 against 415), Larson's method

Transformation method (or image source)	NI	PV	NV
Khan et al.'s TMO [Kha+18]	178	732	870
Ashikhmin et al.'s TMO [Ash02]	178	230	1372
Durand et al.'s TMO [DD02]	138	179	1063
Fattal et al.'s TMO [FLW02]	75	260	415
Reinhard et al.'s TMO [Rei+02]	253	1549	725
Larson et al.'s TMO [LRP97]	127	730	413
Paul et al.'s MEF [PSA16]	97	582	291
Pece et al.'s MEF [PK10]	91	422	377
Raman et al.'s MEF [RC09]	133	945	252
Local Adjustment for MEF	50	57	393
Global Adjustment for MEF	59	409	122
Surreal effect (post processing)	131	117	1062
Grunge effect (post processing)	98	47	835
Flickr dataset	292	1631	997

TABLE 4.2 – The distribution of the naturalness evaluations with respect to each transformation method (or image source) for the whole dataset. NI : number of images, PV : number of positive votes (evaluating images as natural images), NV : number of negative votes (evaluating images as unnatural images).

(730 against 413), Flickr dataset (1631 against 997).

The images are categorized into 10 groups based on the number of positive and negative evaluations they got. The results are showed in Fig. 4.7. A group is represented by a column in the chart. For example, the first left column in the chart corresponds to the 301 images that have been assessed as unnatural by the 9 observers (no one assessed them as natural) while the right last column shows that 143 images have been evaluated as natural by the 9 observers (no one evaluated them as unnatural).

Because the purpose of the research is to study naturalness and unnaturalness signs, there is a need of relevant data. Thus, only the images with a significant difference between the number of positive evaluations and the number of negative evaluations have been considered. Based on the results of the experiment, an image in this study is considered as natural if there are at least 8 positive evaluations (in total 9 evaluations). Similarly, if there are at least 8 negative evaluations (in total 9 evaluations), it is considered as unnatural. The others are considered as uncertain images because related to controversial evaluations.

After discarding the uncertain images, 531 unnatural images and 355 natural images are kept. The details of the evaluation distribution of the reduced version are described in Table 4.3. Obviously, natural images and unnatural images have been generated by different transformation methods. According to the results after discarding uncertain images, it seems that some methods generate mainly natural images (Reinhard's method) or unnatural images (Ashikhmin's method). some methods generate both natural and unnatural images such as

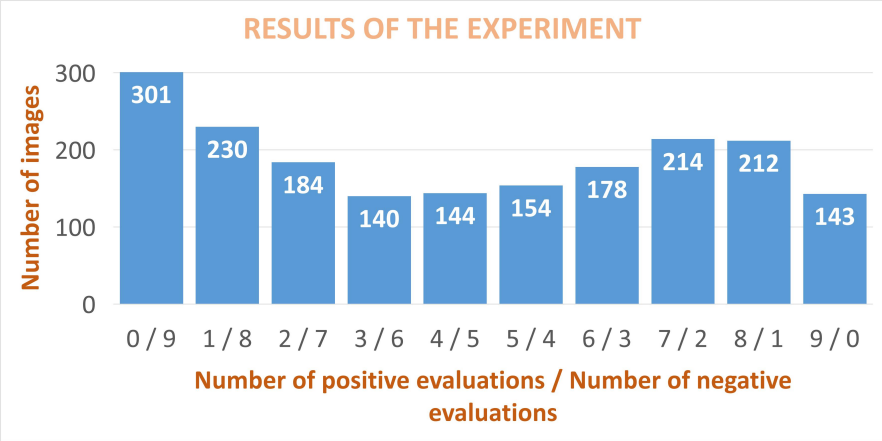


FIGURE 4.7 – Results of the subjective naturalness experiment.

Transformation method (or image source)	NI	PV	NV
Khan et al.'s TMO [Kha+18]	59	231	300
Ashikhmin et al.'s TMO [Ash02]	123	73	1034
Durand et al.'s TMO [DD02]	98	61	821
Fattal et al.'s TMO [FLW02]	21	31	128
Reinhard et al.'s TMO [Rei+02]	93	696	141
Larson et al.'s TMO [LRP97]	45	317	88
Paul et al.'s MEF [PSA16]	33	257	40
Pece et al.'s MEF [PK10]	13	84	33
Raman et al.'s MEF [RC09]	67	560	43
Local Adjustment for MEF	36	18	306
Global Adjustment for MEF	31	265	14
Surreal effect (post processing)	101	28	881
Grunge effect (post processing)	87	18	765
Flickr dataset	79	544	167

TABLE 4.3 – The distribution of the naturalness evaluations for the selected images (images with at least 8 positive votes or 8 negative votes). NI : number of images, PV : number of positive votes (evaluating images as natural images), NV : number of negative votes (evaluating images as unnatural images).

FIGURE 4.8 – Examples of data augmentation including re-scaling, shifting, flipping, cropping and padding (the black padding parts in those image are not presented to the observers). The two first rows present augmented versions of a natural image while the two last rows present augmented versions of an unnatural image (based on observers' evaluations). The data augmentation operations do not change the feeling of naturalness or unnaturalness so that the same label is kept.

Khan's method (231 positive votes versus 300 negative votes), and Pece's method (84 positive votes against 33 negative votes).

In order to balance the dataset 176 unnatural images are removed randomly. Then, the ground-truth of the image naturalness dataset is built from 355 unnatural and 355 natural photos. After applying data augmentation including re-scaling, shifting, flipping, cropping and padding, 200 modified versions of size 224×244 are generated from every original photo (See examples in Fig. 4.8) and the labels of the augmented versions are set the same as the label of the original one. Totally, there are 142,000 images in the naturalness dataset in which half of them are natural and the others are labelled as unnatural.

4.4 Feature definition and feature selection

There is a lot of factors responsible for the unnaturalness of an image. Some of them can be described and defined by looking at the images while it is not easy to explain and modelize the others (see examples in Fig. 4.4). As a consequence, in this study, the considered features for the purpose of INA are built based on the one side on hancrafted features designed to take into account some a priori about unnaturalness and on the other side on features learned directly either from CNNs or from pre-trained models (in order to access to non priori, indescribable information). The proposed handcrafted features are designed to focus on the popular artifacts induced by TMO, MEF and post-processing methods such as the feeling of perceived luminance, contrast, reproduction of detail and colors, bloom, halo and dark band effects [CS05]. In contrast, learned features are used to detect the abstract factors causing an unnaturalness feeling about photos.

4.4.1 Handcrafted features

Based on the ideas mentioned in [CS05] the considered handcrafted features are :

4.4.1.1 Brightness features

SDR images generated by TMO, MEF or post-processing algorithms sometimes look unnatural because of the perceived brightness. The brightness channel is one of the 3 channels of the HSV (or HSB : Hue, Saturation and Value or Brightness) color space. The brightness of a pixel is also calculated as the maximum value of the red, green, blue values (see Eq. 4.7). By analyzing the brightness histogram of the photos in the dataset, some artifact signs related to brightness could be detected. As an example, in Fig. 4.9, according to the results of the experiment, the top left image looks more natural to the observers than the other color images. Looking at the brightness histogram in the last row, it appears that the density of medium brightness values in the natural image seems to be denser than those of the others. In contrast, the two other images look too bright or too dark which can be detected on the

brightness histograms that are distributed more in high or low values. The features representing the brightness histogram including mean (f_1), standard deviation (f_2), skewness (f_3), kurtosis (f_4) and continuity (f_5) of brightness are the first handcrafted features for INA. In which, the continuity of brightness is defined as :

$$f_5 = \sum |H_{br}(i) - H_{br}(i+1)| \tag{4.1}$$

where $H_{br}(i)$ and $H_{br}(i+1)$ are the values of the i^{th} and $i+1^{th}$ bins in the brightness histogram.

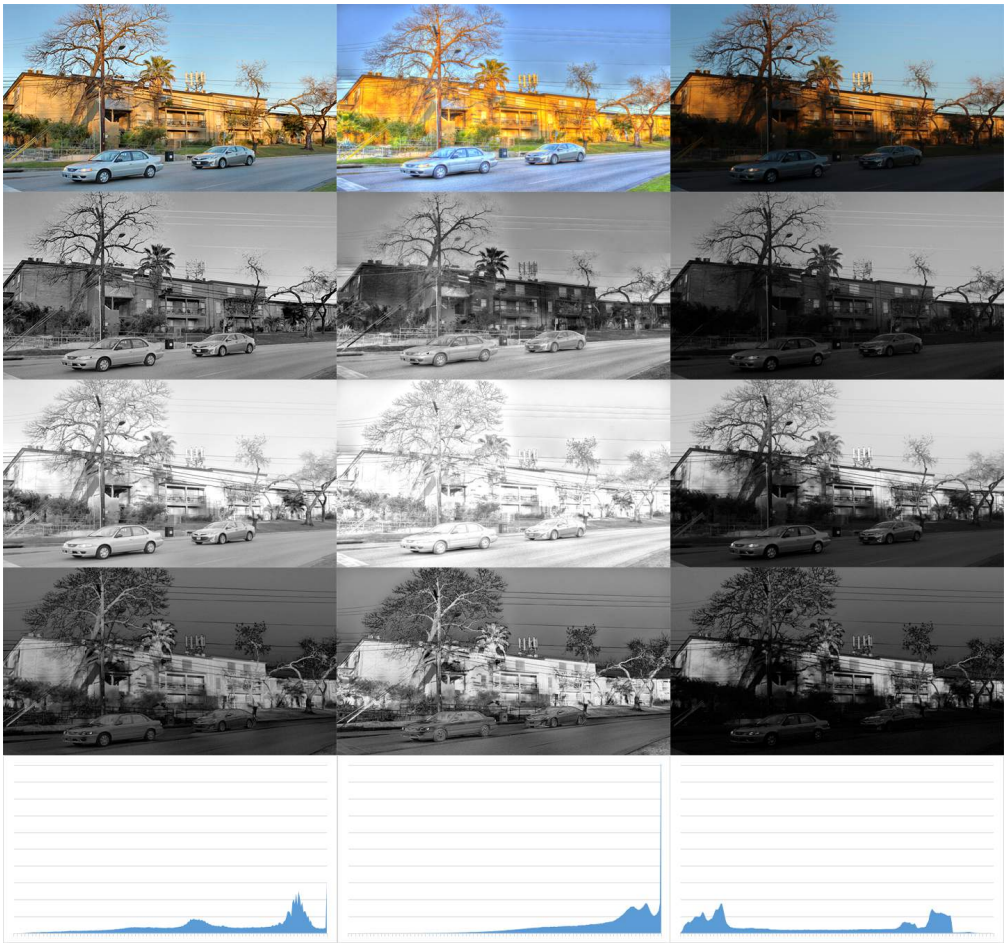


FIGURE 4.9 – The first row presents the color images. The second and the third rows illustrate the corresponding darkness and brightness channels (Eq. 4.6 and Eq. 4.7) of them. The fourth row shows the absolute difference (Eq. 4.8) between the darkness and the brightness channels. The brightness histograms of the color images are presented in the last row.

Another important factor affecting the image naturalness is the brightness contrast. Obviously, the global brightness contrast of an over-exposed image or an under-exposed image is often low. However, a photo with a too high global brightness contrast could also look unnatural. In this study, the features representing the global brightness contrast of an image are

defined as :

$$f_6 = \frac{\mu_{br}^h}{\mu_{br}} \tag{4.2}$$

$$f_7 = \frac{\mu_{br}^l}{\mu_{br}} \tag{4.3}$$

$$f_8 = \frac{f_6 - f_7}{f_6 + f_7} \tag{4.4}$$

where f_6 and f_7 represent the highest and the lowest brightness values normalized by the brightness mean (μ_{br}) respectively. μ_{br}^h and μ_{br}^l are the means of the brightness values of top 5 percent pixels having the highest and the lowest brightness values respectively. The global brightness contrast (f_8) represents the relation between the highest and lowest brightness values in the photo. An unnatural photo often has a too high or a too low contrast or it also could have low highest brightness values (under-exposed images) or high lowest brightness values (over-exposed images).

FIGURE 4.10 – The left column presents the color images. The right one illustrates the corresponding brightness channels of them. The first left image labelled as unnatural contains artifact signs : halo, dark band and bloom effects while the second color image is assessed as natural by the observers.

Additionally, images mapped by some TMOs have artifact signs such as dark bands, halos and bloom effect (see examples in Fig. 4.1 and Fig. 4.10). Halos and dark bands surrounding details increase the contrast of local parts as in Fig. 4.10 (halos have high brightness values while dark bands have low brightness values). In order to detect the high contrast of local parts caused by halo and dark band effects, an image is divided into M parts and the local brightness contrast of the image is defined as the mean of brightness contrasts of the M parts :

$$f_9 = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{br}^i - min_{br}^i}{max_{br}^i + min_{br}^i} \tag{4.5}$$

where max_{br}^{i} and min_{br}^{i} are the maximum and the minimum brightness values respectively in the i^{th} part. In this work, M is set to 100 (10×10 as in the right column of Fig. 4.10).

4.4.1.2 Saturation features

The impression of colors is not only caused by brightness factors but is also affected by saturation factors. Thus saturation factors have significant influences on naturalness perception of images. Similarly to the brightness features, 9 features (f_{10} to f_{18}) are defined based on saturation information (extracted from the channels of the HSV color space) to present the saturation distribution and the saturation contrast of an image.

4.4.1.3 The darkness channel and its relation with the brightness channel

Analyzing the darkness channel and its relation with the brightness channel is an effective way to classify over-exposed, under-exposed and well exposed images. Considering an image in the RGB color space, the darkness channel (I_{da}) and the brightness channel (I_{br}) are defined based on the RGB channels as :

$$I_{da}(x,y) = \min\big(R(x,y), G(x,y), B(x,y)\big) \tag{4.6}$$

$$I_{br}(x,y) = \max\big(R(x,y), G(x,y), B(x,y)\big) \tag{4.7}$$

where (x,y) are the coordinates of a pixel. $R(x,y)$, $G(x,y)$ and $B(x,y)$ are red, green and blue levels at point (x,y) respectively. The difference between the two channels is defined as :

$$I_{di}(x,y) = I_{br}(x,y) - I_{da}(x,y) \tag{4.8}$$

In Fig. 4.9, it appears that the darkness values of the under-exposed image in the last column are very low while the brightness values of the over-exposed image in the second column are too high. Beside this, the difference between the brightness and darkness channels of the over-exposed image is higher than that of the well exposed image. In contrast, this difference for the under-exposed image is less significant than that of the natural one. Therefore the information of the darkness channel and its relation with the brightness channel is an important clue to evaluate the naturalness of a photo. The 8 next features (f_{19} to f_{26}) for INA are the mean, standard deviation, kurtosis, skewness of I_{da} and I_{di} respectively.

Obviously, some details in the darkness and brightness channels of the images in the 2 last columns (Fig. 4.9) are lost. By comparing the details of the original image in gray scale and the details of the darkness and brightness channels, the reproduction of details and the balance between the darkness and the brightness channels can be evaluated. Thus the 2 last handcrafted features are defined as :

$$f_{27} = \frac{\sum |G_{I_g} - G_{I_{da}}|}{\sum G_{I_g}} \tag{4.9}$$

$$f_{28} = \frac{\sum |G_{I_g} - G_{I_{br}}|}{\sum G_{I_g}} \tag{4.10}$$

where $G_{I_g}, G_{I_{da}}, G_{I_{br}}$ are the gradient images [ADG12] of the original image in gray scale, the darkness and the brightness channels respectively. $\sum G$ is the sum of pixel values of the image G. Note that the black padding regions (generated by the data augmentation methods) of images are discarded before calculating the handcrafted features.

To sum up, the overview of the considered handcrafted features is presented in Table 4.4. The features are categorized in 3 groups including brightness features (9 features), saturation features (9 features) and darkness features (10 features).

4.4.2 Learned features

In some cases, it is possible to explain why an image looks unnatural to an observer but in general, it is a tough task. No direct relation appears between the unnatural feeling and the image clues such as color, brightness, saturation and so on. As a result, besides being handcrafted, features have also to be learned directly from images by using CNNs [KSH17]. Because of the modest image number of the naturalness dataset, the 2 approaches used for learning features in this study are shallow CNNs and transfer learning [PY10] (using deep features learned from pre-trained deep CNNs).

4.4.2.1 Shallow learned features

In the first approach, shallow learned features are learned from shallow CNNs. Shallow CNNs are defined in this study as models with a low number of convolutional layers and a shallow architecture. In this work, a very shallow architecture with only one convolutional layer is chosen. The general structure of the 4 considered models (see Fig. 4.11) includes a convolutional layer receiving input color images of size 224×224, a global average pooling layer transforming n-D outputs from the convolutional layer into 1D outputs, a batch normalization layer normalizing the outputs from the global pooling layer and a fully connected layer on the top for predicting the input images as natural or unnatural. The size and the number of kernels in the convolutional layer are designed according to the number of samples in the dataset (142,000 samples of size 224×224). In order to learn various types of features, different models using different kernel sizes and different kernel numbers (490 kernels of size 5×5, 229 kernels of size 9×9, 65 kernels of size 17×17 and 65 kernels of size (2×17)×(2×17) - an average pooling layer is used to resize the input image by 50 percent) are designed as in Fig. 4.11. After the training phase, the models without the prediction layer are considered as feature extractors computing the learned features from the input images. The 4 feature extractors calculate 65, 65, 229 and 490 shallow learned features (features learned from shallow CNNs) for the purpose of INA. In the training process, the Adam optimizer and a binary cross-entropy loss function are used and the batch size is assigned to 128. The learning rate and the number of iterations are set to 10^{-6} and 3000 respectively.

Features	**Formula**
Brightness features	$f_1 = \frac{\sum_{i=1}^{N} I_{br}(i)}{N}$ $f_2 = \sqrt{\frac{\sum_{i=1}^{N} (I_{br}(i)-f_1)^2}{N-1}}$ $f_3 = \frac{\sum_{i=1}^{N} (I_{br}(i)-f_1)^3}{N \times f_2^3}$ $f_4 = \frac{\sum_{i=1}^{N} (I_{br}(i)-f_1)^4}{N \times f_2^4}$ $f_5 = \sum \lvert H_{br}(i) - H_{br}(i+1) \rvert$ $f_6 = \frac{\mu_{br}^h}{\mu_{br}}$ $f_7 = \frac{\mu_{br}^l}{\mu_{br}}$ $f_8 = \frac{f_6 - f_7}{f_6 + f_7}$ $f_9 = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{br}^i - min_{br}^i}{max_{br}^i + min_{br}^i}$ I_{br} is the brightness channel.
Saturation features	$f_{10} = \frac{\sum_{i=1}^{N} I_{sa}(i)}{N}$ $f_{11} = \sqrt{\frac{\sum_{i=1}^{N} (I_{sa}(i)-f_{10})^2}{N-1}}$ $f_{12} = \frac{\sum_{i=1}^{N} (I_{sa}(i)-f_{10})^3}{N \times f_{11}^3}$ $f_{13} = \frac{\sum_{i=1}^{N} (I_{sa}(i)-f_{10})^4}{N \times f_{11}^4}$ $f_{14} = \sum \lvert H_{sa}(i) - H_{sa}(i+1) \rvert$ $f_{15} = \frac{\mu_{sa}^h}{\mu_{sa}}$ $f_{16} = \frac{\mu_{sa}^l}{\mu_{sa}}$ $f_{17} = \frac{f_{15} - f_{16}}{f_{15} + f_{16}}$ $f_{18} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{sa}^i - min_{sa}^i}{max_{sa}^i + min_{sa}^i}$ I_{sa} is the saturation channel.
Darkness features	$f_{19} = \frac{\sum_{i=1}^{N} I_{da}(i)}{N}$ $f_{20} = \sqrt{\frac{\sum_{i=1}^{N} (I_{da}(i)-f_{19})^2}{N-1}}$ $f_{21} = \frac{\sum_{i=1}^{N} (I_{da}(i)-f_{19})^3}{N \times f_{20}^3}$ $f_{22} = \frac{\sum_{i=1}^{N} (I_{da}(i)-f_{19})^4}{N \times f_{20}^4}$ $f_{23} = \frac{\sum_{i=1}^{N} I_{di}(i)}{N}$ $f_{24} = \sqrt{\frac{\sum_{i=1}^{N} (I_{di}(i)-f_{23})^2}{N-1}}$ $f_{25} = \frac{\sum_{i=1}^{N} (I_{di}(i)-f_{23})^3}{N \times f_{24}^3}$ $f_{26} = \frac{\sum_{i=1}^{N} (I_{di}(i)-f_{23})^4}{N \times f_{24}^4}$ $f_{27} = \frac{\sum \lvert G_{I_g} - G_{I_{da}} \rvert}{\sum G_{I_g}}$ $f_{28} = \frac{\sum \lvert G_{I_g} - G_{I_{br}} \rvert}{\sum G_{I_g}}$ I_{da} is the darkness channel. I_{di} is the differences between the brightness and darkness channels. $I_{di} = I_{br} - I_{da}$

TABLE 4.4 – Overview of the proposed handcrafted features for INA.

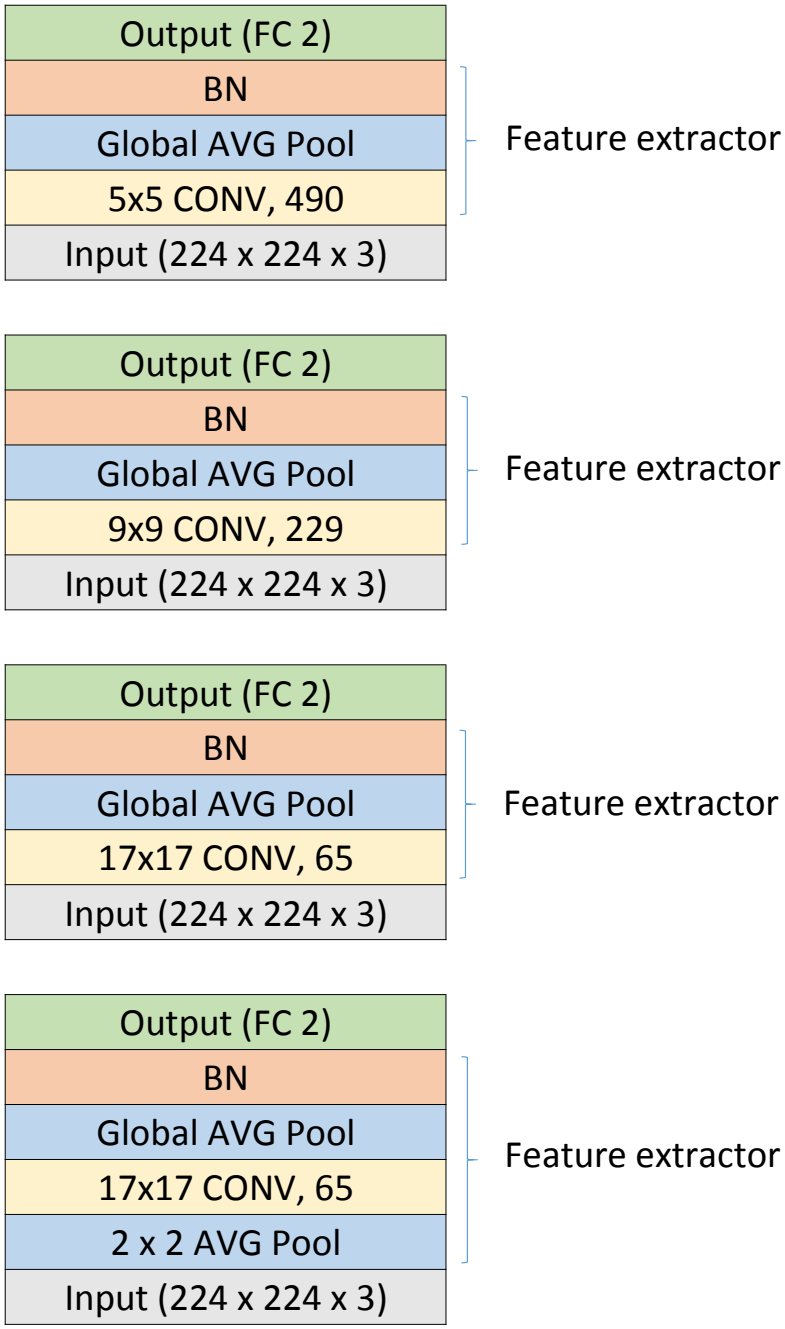


FIGURE 4.11 – Four different architectures of the shallow CNN. 2×2 AVG Pool : Average pooling layer with the pooling of size 2×2 that reduces the size of the input image by 50 percent. W×W CONV, N : N kernels of size W×W of the convolutional layer. Global AVG Pool : global average pooling layer. BN : Batch normalization layer. FC 2 : The fully connected layer containing 2 output neurons (the prediction layer).

4.4.2.2 Deep learned features

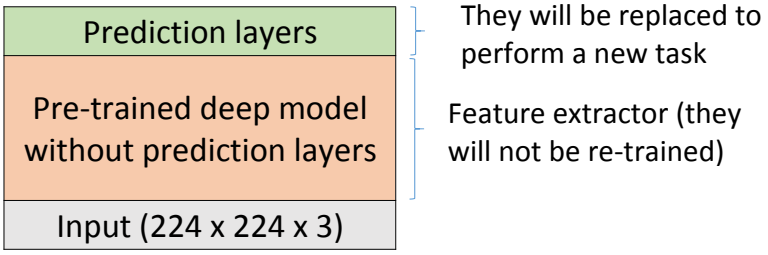


FIGURE 4.12 – The general architecture of the transfer deep models.

Deep learned features are learned from deep CNNs (models with a high number of convolutional layers and a deep architecture). To learn deep features directly, there is a need of a very high number of images and it is impossible in the case of this study. Considering deep features learned by pre-trained models could be a good solution. Although the deep learned features (features learned from deep CNNs) have been learned for a given task, they can be considered to be used for different tasks [PY10]. The general structure of deep CNNs includes convolution layers at the bottom and fully connected layers on the top. The convolution layers are responsible for learning features while the fully connected layers are in charge of combining features learned from the convolution layers to solve the task. In other words, after removing the fully connected layers of a pre-trained deep neural network, the model can be considered as a feature extractor. In this study, several deep models including VGG16 [SZ15], Xception [Cho16], ResNet [He+15], NASNet large and NASNet mobile [Zop+17], MobileNet [How+17], Inception [Sze+15], DenseNet [HLW16], Inception ResNet [SIV16] pre-trained on the ImageNet dataset for the task of image classification are transferred to the new purpose of INA by keeping the convolution layers and replacing the top fully connected layers and training them for the new task. The general structure of the transfer models is presented in Fig 4.12. Additionally, instead of using all pre-learned features for the new task, there is a feature selection process to extract the most relevant features.

When using transfer learning, features are primarily learned for a different task. Some features can be transferred well to perform a new task while the remaining are not relevant. Additionally, combining several feature sets could increase the performance but it also increases the number of features. This increase makes the computation complicated and sometimes it also increases the requirement of data. Therefore, simplifying a feature set by selecting the most relevant features is a good solution. The feature reduction algorithm mentioned in 2.3 is then applied on the initial feature set to select the most relevant features to perform the task. After selecting the most relevant features, the accuracy of the classification based on shallow learned features increases from 0.789 (with 849 features) to 0.808 (with 731 features) while the accuracy of the classification based on the deep learned features improves from 0.858 (with 2048 features) to 0.865 (with 425 features) respectively. The feature reduction helps first and foremost reducing the number of features to be considered.

4.5 Experiments and results

There are 2 main purposes in this part. The first goal is to automatically answer the question "does an image look natural or not ?". This is dealt via a binary classification approach. The second one is to look for the most efficient features for INA.

INA is considered here as a binary classification problem (natural / unnatural image) reflecting the fact that an observer might or might not feel that the image is natural. In order to evaluate the performance of each feature set, the classification is performed separately with the handcrafted features, the shallow learned features and the deep learned features and the combinations of them.

4.5.1 Dataset and setup

The general model having the general structure as in Fig. 4.13 is trained and tested to evaluate the classification performances of each feature set. The structure includes an input layer, an output layer and P hidden blocks (in this study, P is set to 4). Each block contains a fully connected layer, a batch normalization layer and a dropout layer. The output layer contains 2 neurons corresponding to the 2 classes (natural and unnatural). The model is designed to learn how to combine the computed features for the classification task. Only the fully connected layers are trained in the training process so the convergence is fast. In the experiment, the number of iterations is set to 150. The Adam optimizer is used and the loss function is the binary cross-entropy loss. The learning rate and the mini-batch size are set to 10^{-3}, 512 respectively. Regarding the feature extraction block, each of the 3 feature sets is tested alone. An SVM classifier is not used in this case because the classification is going to be trained on over 100,000 samples (it will take months to train the SVM classifier with these samples).

The model is trained on the 113,600 images of the training set S (56,800 natural images and 56,800 unnatural images coming from 284 original natural images and 284 original unnatural images) and tested on 2 testing sets : S'_1 including 28,400 images (14,200 natural images and 14,200 unnatural images) generated from the 142 remaining original images by applying the data augmentation process and S'_2 containing 142 images (71 natural images and 71 unnatural images) obtained from 142 original images by re-scaling and padding (just to convert images to the format of size 224×224 without cropping). The classifier is tested on the 2 testing sets to evaluate the influence of data augmentation on performances. It helps to demonstrate the validity of the data augmentation process regarding the labeling in particular. There is no overlapping images (images generated from the same original images) between the training set and the testing set.

As before, the model is evaluated based on the Accuracy (A). In general, the accuracy (or overall accuracy) is the most popular metric for evaluating classification performance while the loss (or mean absolute error) reflects the classification certainty. In Table 4.5, n is the number of classes (in this case $n = 2$), y and o are the target and the output (prediction)

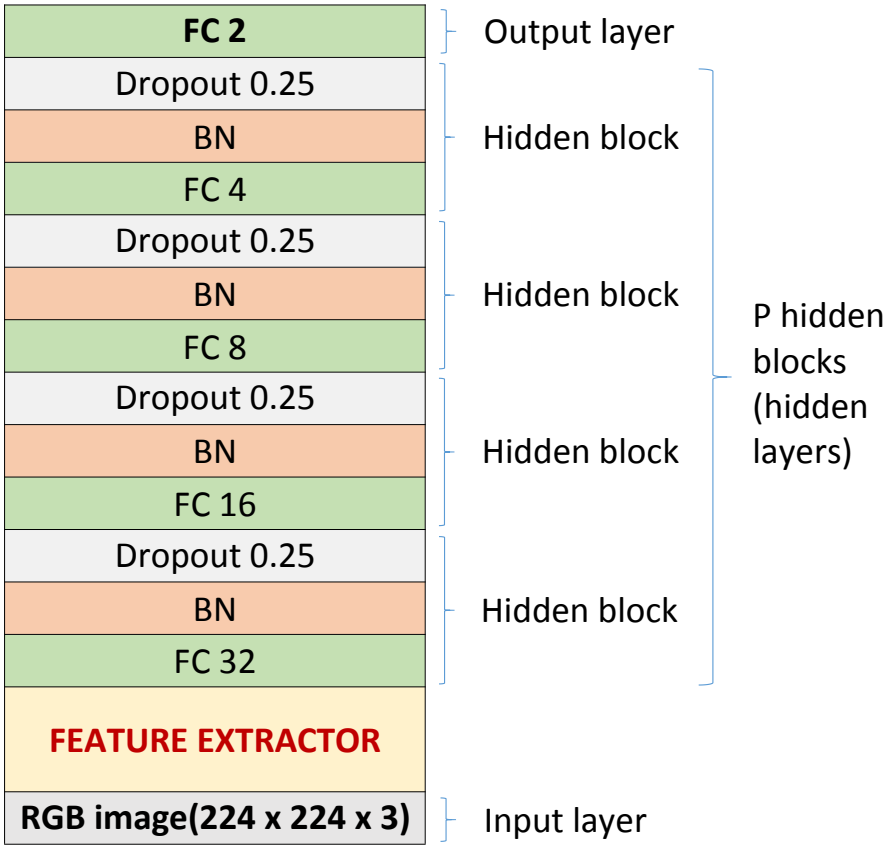


FIGURE 4.13 – General structure of the network designed for natural / unnatural image classification. Features extracted from an RGB input image of size 224×224×3 by the feature extractor are passed through the layers to classify the image as natural or unnatural. There are 4 hidden blocks with a fully connected layer, a batch normalization layer and a dropout layer in each block.

Evaluation criteria	Formula
Accuracy	$A = \frac{TP+TN}{TP+FP+TN+FN}$
Lower accuracy	$A_l = A - I_a$
Upper accuracy	$A_u = A + I_a$
Loss	$L = \frac{\sum_{i=1}^{n} \lvert y_i - o_i \rvert}{n}$
Lower loss	$L_l = L - I_l$
Upper loss	$L_u = L + I_l$

TABLE 4.5 – Overview of evaluation criteria for INA.

values respectively. The lower loss (L_l) and the upper loss (L_u) present the range of loss within the 95% confidence interval [Mit97] ; [DE96]. I_a and I_l are the accuracy interval and the loss interval.

The experiments have been performed on a PC equipped with an Intel(R) Xeon(R) CPU X5650 2.67 GHz (12 CPUs) and 24 GB memory.

4.5.2 Results and discussions

4.5.2.1 Handcrafted features based classification

In this first case, the feature extractor block in Fig. 4.13 computes the handcrafted features defined in section 4.4.1. Table 4.6 shows the performances of the classification based on handcrafted features. The impact of each handcrafted feature subset (brightness features, saturation features and darkness channel features) is also estimated and is showed in the table. It appears that the overall accuracy of classification based on the separate feature subsets is quite low (0.712, 0.640, 0.716 for the brightness, saturation and darkness features respectively) and the loss is high (0.389, 0.495, 0.401 for the brightness, saturation and darkness features respectively). By combining them, the overall accuracy increases to 0.812 and the loss decreases to 0.321. Beside this, it appears that the FP value is much higher than the FN value (3832 versus 1520), the handcrafted features appear to be less sensitive to unnatural images in this case because of the limitation of handcrafted features.

Classification examples based on handcrafted features are shown in Fig. 4.14. As expected, the unnatural images caused by low saturation are well classified with those features. Although halos around details and contrast factors have been considered during the feature design stage, there are some images with those artifacts in the misclassified unnatural images. Additionally, most of the well classified unnatural images and the misclassified natural images are colorful while the well classified natural images and the misclassified unnatural images are less colorful. Clues regarding colorfulness are one typical unnaturalness signs so some handcrafted features in the proposed feature set have been designed to detect this kind of clues. It explains why the classifier based on the handcrafted features appears to be too sensitive to colorfulness. It seems that the handcrafted features are not able to detect all the cases and sometimes they

Classification based on handcrafted feature subsets performed on the testing set S'_1		
9 Brightness features	A = 0.712	L = 0.389
9 Saturation features	A = 0.640	L = 0.495
10 Darkness channel features	A = 0.716	L = 0.401

Classification based on all the handcrafted features (28 features) performed on the testing set S'_1			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,680	FN = 1520
	Unnatural	FP = 3832	TN = 10,368
A = 0.812	I_a = 0.005	A_l = 0.807	A_u = 0.817
L = 0.321	I_l = 0.005	L_l = 0.316	L_u = 0.326

TABLE 4.6 – INA based on the 28 handcrafted features and impact of each handcrafted feature group on the assessment.

are too sensitive to some factors. So some discriminant features are not taken into account with the considered handcrafted features.

4.5.2.2 Shallow learned features based classification

In Fig. 4.13, the feature extractors are now made of the shallow CNNs described in section 4.4.2.1. Beside the classifications based on separate feature sets, the classification with the combination of all the shallow learned features is also performed. The details of classification using features learned from the 4 shallow CNNs are shown in Table 4.7. Obviously, the classification based on the combination of shallow learned features has the best overall accuracy (0.786) and the best loss (0.269) but the number of features is also the highest (849 features) among the shallow learned feature sets (65, 65, 229 and 490 features).

In order to study the compromise between the number of features and the accuracy, the feature reduction algorithm based on the Relief method presented in Chapter 2 is applied on the combined feature set to reduce the feature number from 849 to 731. Although the number of features decreases, the feature computational time does not change because the two feature sets are computed by the same CNNs. By keeping the most relevant features only for the classification, the overall classification accuracy increases slightly from 0.786 to 0.808 while the loss values are almost the same (0.269 and 0.274).

Fig. 4.15 shows classification examples based on the combination of the shallow learned feature sets. Focusing on the true classification samples of unnatural images, it appears that the filters in the shallow models are efficient to detect unnatural images caused by halos around details. Contrary to the classification based on handcrafted features, the shallow learned features based classifier is not efficient to detect color saturation artifacts since the color

FIGURE 4.14 – Classification examples with handcrafted features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated to a very low loss value while the two right columns contain misclassified images associated to a very high loss value.

saturation of 5 (of the 8) misclassifed unnatural images is low. With the handcrafted features, the classifier focuses on the characteristics of the whole image while the shallow learned features based classifier focuses on each sub region of the image (the size of sub regions depends on the size of kernels). It explains the differences between the classification results based on the two feature sets.

4.5.2.3 Deep learned features based classification

The feature extractor is successively made of the nine pre-trained deep models described in section 4.4.2.2 followed by the feature selection process described in section **??**. After training and testing the models using the 9 reduced feature sets, the highest overall accuracy (0.865) and the best loss (0.139) are obtained with the model using the features learned from the ResNet extractor (see Table 4.8). The ResNet model was pre-trained on ImageNet dataset using an SGD optimizer, a batch size of 256, a momentum of 0.9. It was trained for 60×10^4 iterations with the learning rate starting at 0.1 and divided by 10 when the error reaches a plateaus [He+15]. In this case, there is no re-trained ResNet layers. The model without the last layer (the fully connected layer) is considered as the feature extractor for the proposed model as in Fig. 4.13. Specifically, 425 learned features are selected from the 2048 ResNet features by applying the Relief based feature reduction algorithm. The details of the best classification are showed in Table 4.9. The overall accuracy and the loss of the classification are quite good at 0.865 and 0.139 respectively.

Classification examples based on the ResNet features are presented in Fig. 4.16. It appears that some of the well classified unnatural images have halos around details. Secondly, brightness factors are not detected well since there are some misclassified unnatural images having a too low brightness. Beside this, it is similar to the handcrafted features based classification since most of the well classified unnatural images are colorful. There are some overlapping images (4 of 8) between the misclassified natural images based on the shallow learned features and the ones based on the deep learned features. It demonstrates that some similar characteristics are learned from the training samples by both deep and shallow CNNs.

4.5.2.4 Discussion about the data augmentation process

The general purpose of the study is the naturalness of images (not naturalness of scenes). There are several images of the same scene generated from the same original image but in various ways and they might look totally different (See the first row of Fig. 4.9 where 3 images of the same scene are generated in 3 different ways). Table 4.10 reflects that the classifications performed on S'_1 and S'_2 are similar since the differences in classification accuracy and classification loss are insignificant. This means that the INA is not affected by the data augmentation process. That is why the same natural or unnatural label after data augmentation is kept. The intervals of accuracy and loss depend on the number of testing samples. Indeed, the accuracy and loss intervals of the classification performed on S'_1 (from 0.004 to 0.005) is much smaller than those of the classification executed on S'_2 (from 0.052 to 0.077) because the number of

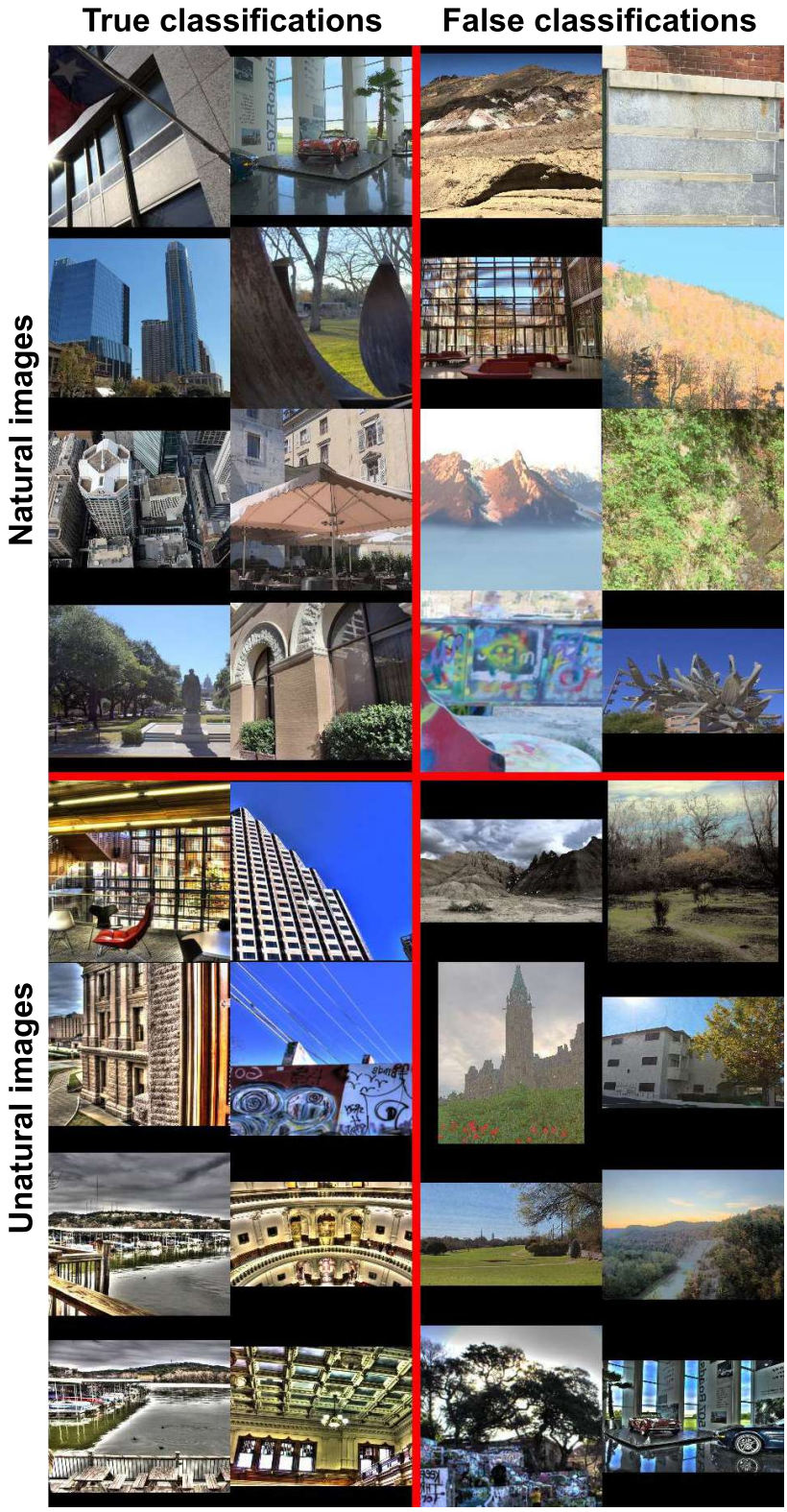


FIGURE 4.15 – Classification examples with shallow learned features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated with a low loss value while the two right columns contain misclassified images associated with a high loss value.

Classification based on the separate feature subsets learned from the shallow convolutional networks performed on the testing set S'_1		
Features learned from the model with 490 5×5 kernels	A = 0.756	L = 0.337
Features learned from the model with 299 9×9 kernels	A = 0.766	L = 0.305
Features learned from the model with 65 17×17 kernels	A = 0.753	L = 0.329
Features learned from the model with 65 17×17 kernels and an average pooling layer	A = 0.741	L = 0.332

Classification based on all the shallow learned features (849 features) performed on the testing set S'_1			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 11,504	FN = 2,669
	Unnatural	FP = 3,376	TN = 10,824
A = 0.786	I_a = 0.005	A_l = 0.781	A_u = 0.791
L = 0.269	I_l = 0.005	L_l = 0.264	L_u = 0.274

Classification based on the reduced shallow learned feature set (731 features) performed on the testing set S'_1			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 11,131	FN = 3,069
	Unnatural	FP = 2,390	TN = 11,810
A = 0.808	I_a = 0.005	A_l = 0.803	A_u = 0.813
L = 0.274	I_l = 0.005	L_l = 0.269	L_u = 0.279

TABLE 4.7 – INA based on the shallow learned features and impact of each shallow learned feature group on the assessment.

Classification based on deep features learned from different deep models performed on the testing set S'_1		
Features learned from	Accuracy	Loss
Xception model	0.678	0.351
NASNet Large model	0.698	0.309
NASNet Mobile model	0.709	0.329
MobileNet model	0.732	0.329
Inception ResNet model	0.729	0.279
Inception model	0.736	0.271
VGG16 model	0.773	0.262
DenseNet model	0.786	0.222
ResNet model	0.865	0.139

TABLE 4.8 – INA based on deep features learned from different deep models.

Classification based on the ResNet feature set (2048 features) performed on the testing set S'_1			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,539	FN = 1661
	Unnatural	FP = 2359	TN = 11,841
A = 0.858	$I_a = 0.004$	$A_l = 0.854$	$A_u = 0.862$
L = 0.299	$I_l = 0.005$	$L_l = 0.294$	$L_u = 0.304$

Classification based on the reduced ResNet feature set (425 features) performed on the testing set S'_1			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,709	FN = 1491
	Unnatural	FP = 2336	TN = 11,864
A = 0.865	$I_a = 0.004$	$A_l = 0.861$	$A_u = 0.869$
L = 0.139	$I_l = 0.004$	$L_l = 0.135$	$L_u = 0.143$

TABLE 4.9 – INA based on the features learned from the ResNet model pre-trained on the ImageNet dataset.

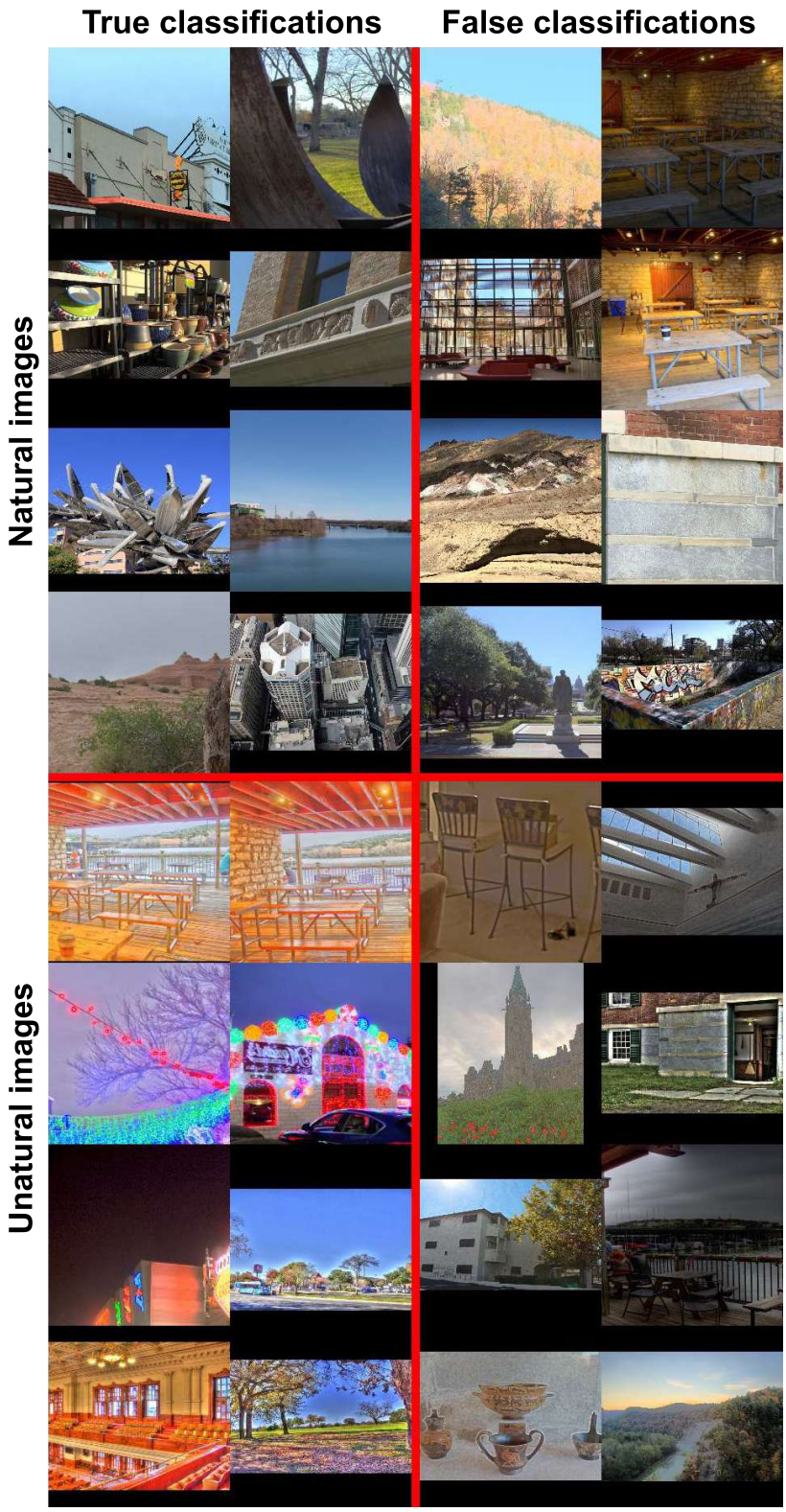


FIGURE 4.16 – Classification examples with deep learned features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated with a low loss value while the two right columns contain misclassified images associated with a high loss value.

Feature set	$A \pm I_a$ (testing on S'_1)	$L \pm I_l$ (testing on S'_1)	$A \pm I_a$ (testing on S'_2)	$L \pm I_l$ (testing on S'_2)
Handcrafted features	0.812 ± 0.005	0.321 ± 0.005	0.831 ± 0.062	0.326 ± 0.077
Shallow learned features	0.808 ± 0.005	0.274 ± 0.005	0.824 ± 0.063	0.277 ± 0.074
Deep learned features	0.865 ± 0.004	0.139 ± 0.004	0.887 ± 0.052	0.132 ± 0.056

TABLE 4.10 – INA based on the 3 feature sets performed on the testing sets S'_1 and S'_2.

samples in S'_1 is much bigger than that of S'_2 (28,400 versus 142). The highest confidence of the classification results is obtained with the biggest testing set S'_1.

4.5.2.5 Discussion about the loss function evolution

In Fig. 4.17, it is seen that the classifications based on different feature sets act differently since some images are classified well with a feature set but they are misclassified with the others. Looking at Fig. 4.18, it appears that the evolution of the loss function is quite different between the feature sets. With handcrafted features and shallow learned features, the loss values constantly increase from low values to high values (even not reaching 0 or 1 in the case of handcrafted features). In contrast, with deep learned features, the loss values in true classifications are nearly zero and they are nearly one in false classifications. This makes the decision more reliable. Moreover, the average loss of the classification based on deep learned features is much smaller than that of the others.

4.5.2.6 General comparison between classifications

Table 4.12 presents a general comparison between the classifications based on the 3 feature sets. By using deep learned features, the classification accuracy and the classification loss reach the best values (Accuracy : 0.865 compared to 0.812, 0.808 and Loss : 0.139 versus 0.321, 0.274 for the handcrafted, shallow learned features based classifications respectively). Additionally, the classification accuracy with the handcrafted features and the one with the shallow learned features are almost equal. The handcrafted features (when well designed) are quite efficient since an overall accuracy of 0.812 is obtained with only 28 features. The classification accuracy with handcrafted features is similar to the accuracy with shallow learned features (accuracy : 0.812 versus 0.808) but the classification certainty for shallow learned features is better (loss : 0.274 against 0.321). Although the number of handcrafted features is smaller than that of the learned features (28 versus 731, 384 and 425), the performance with handcrafted features is competitive. Deep learned features are quite efficient since the classification accuracy and the

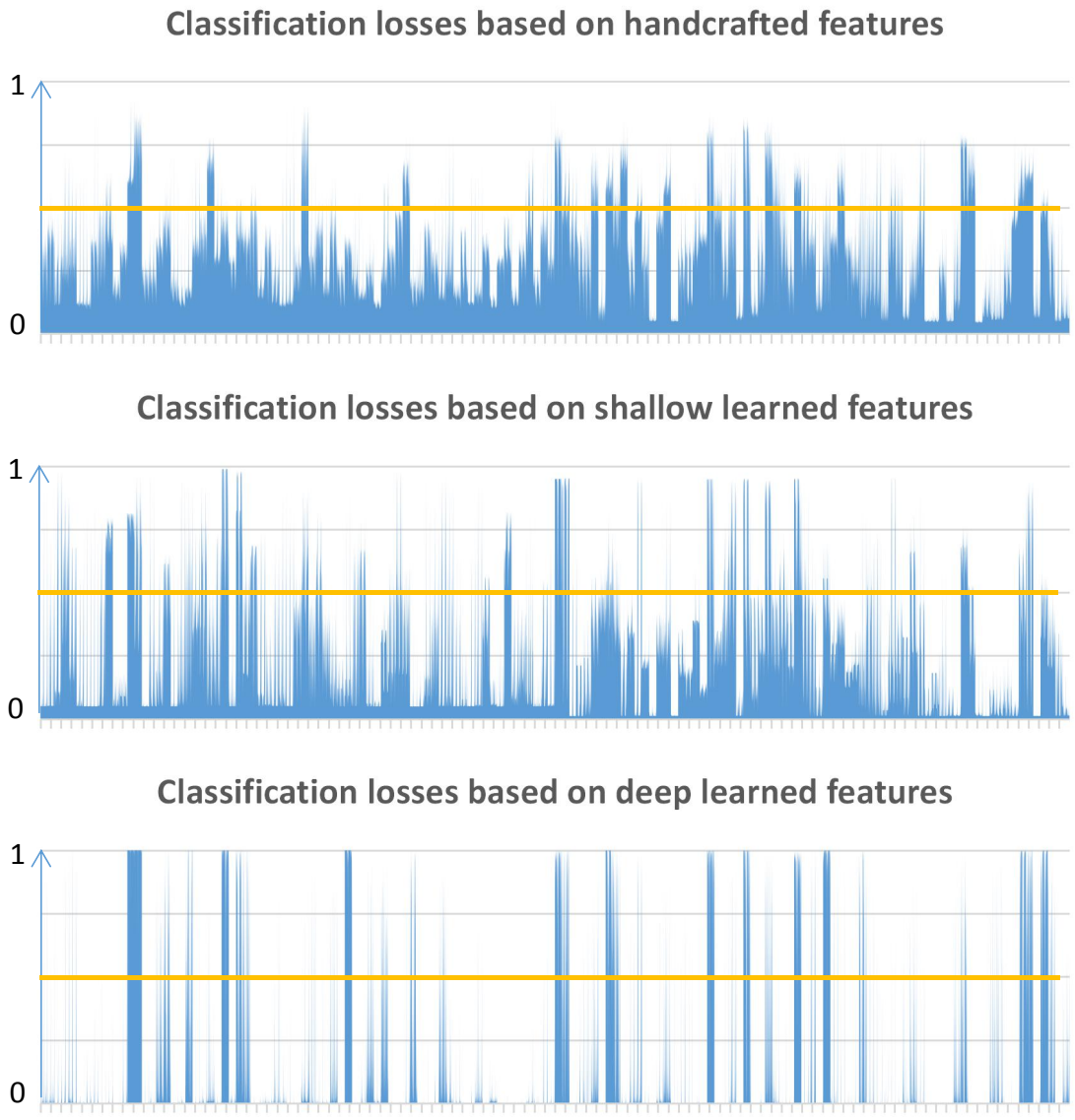


FIGURE 4.17 – Classification losses based on the 3 feature sets. Y axis represents the loss values while X axis represents the images. Each horizontal line is the border between true classifications (loss < 0.5) and false classifications (loss > 0.5).

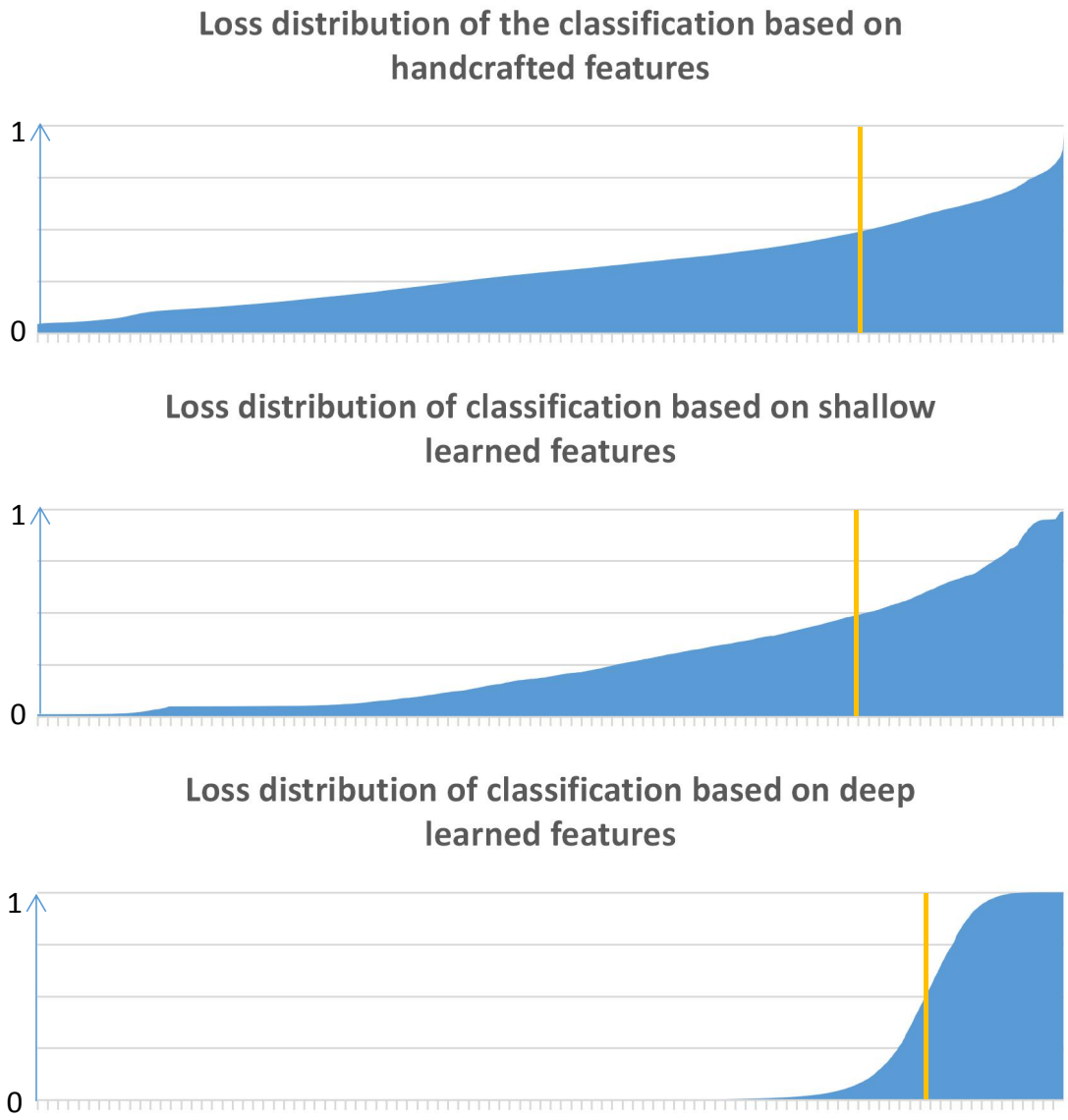


FIGURE 4.18 – Loss distribution of classification based on the 3 feature sets. Y axis represents the loss values while X axis represents the images (sorted based on loss values). Each vertical line is the border between true classifications (loss < 0.5) and false classifications (loss > 0.5).

		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 59,165	FN = 11,835
	Unnatural	FP = 24,190	TN = 82,010
A = 0.797		L = 0.206	

TABLE 4.11 – Cross validation of the model using the reduced ResNet feature set (425 features). Each group of images is considered as the testing set while the remaining groups are considered as the training set.

loss are better than those of handcrafted features and shallow learned features (accuracy : 0.865, loss : 0.139). It is seen that the problem of naturalness is abstract and too complicated for shallow CNN architectures to learn features reflecting this problem. Using simple and shallow CNN architectures could not be a good choice for this problem (the classification performance with shallow learned features is even lower than that with handcrafted features).

4.5.2.7 Transformation signature learning ?

The last discussion of this part is "did the models really learn to access naturalness / unnaturalness or did they just learn to recognize the signature of the corresponding transformation methods ?" because unnaturalness signs come from transformation methods. Looking at Table 4.3, it appears that different sources of images have been considered. And negative and positive evaluations are not coming all from the same source of images. Additionally, there are few images per group, so it is unlikely that the signatures of the transformation methods are learned in this case. In order to verify this assumption, an additional experiment has been performed. The images are categorized in 14 groups as in Table 4.3. The experiment is performed 14 times, each time the augmented images coming from only one group are considered as the testing set while the classifier using the reduced ResNet feature set (425 features) is trained with the augmented images from the remaining ones, so the signatures of the transformation method in the testing set can not be learned. The initialization of the training process is similar to that of the previous one with the reduced ResNet feature set. The results are showed in Table 4.11. Although the accuracy and the loss change a little bit compared to the accuracy and the loss of the model trained in the previous way (from 0.865 to 0.797 for accuracy and from 0.139 to 0.206 for loss), the accuracy and the loss values are quite good at 0.797 and 0.206 respectively and the differences are in-significant. Therefore, it can be concluded that the extracted features are for INA and the solved problem here is definitely not the classification of the transformation methods.

Feature set	Number of features	$A \pm I_a$	$L \pm I_l$
Handcrafted features	28	0.812 ± 0.005	0.321 ± 0.005
Shallow learned features	731	0.808 ± 0.005	0.274 ± 0.005
Deep learned features	425	0.865 ± 0.004	0.139 ± 0.004
Combination of handcrafted and shallow learned features	759	0.827 ± 0.005	0.221 ± 0.005
Combination of handcrafted and deep learned features	453	0.873 ± 0.004	0.134 ± 0.004

TABLE 4.12 – INA performed on the testing set S_1' based on the handcrafted, the shallow and the deep learned features and the combinations of the handcrafted features and learned features.

4.5.2.8 Combining the different feature sets ?

Looking at Fig. 4.14, Fig. 4.15 and Fig. 4.16, there are 9 overlapping samples between classification samples for shallow and deep learned features so there might be some similarities between the 2 feature types. In contrast, the samples for handcrafted features are almost different from the samples for shallow and deep learned features meaning that learned features and handcrafted features might be complementary. In order to validate the assumption, classification performances based on the combinations of handcrafted features and learned features are presented in the 2 last rows of Table 4.12. Obviously, the combination of features it helps improving the classification accuracy from 0.808 and 0.865 to 0.827 and 0.873 for shallow learned features and deep learned features respectively. It proves that learned features and handcrafted features exploit different aspects of image naturalness.

4.5.2.9 INA on lower confidence data

In this study, the classifier based on the ResNet features works quite well on the high confidence data (data whose labels have been decided by at least 8 of 9 observers). But "how will the classifier work on lower confidence data ?". An additional experiment is performed with the classifier based on the ResNet features and the lower confidence data coming from the results of the experiment in part 4.3 to see the differences between the classifications on high confidence data and on lower confidence data. Three image groups containing images with the naturalness labels decided by 7, 6, 5 of 9 observers are extracted from the image set and are augmented (200 augmented versions generated from 1 original image by re-scaling, cropping, padding, flipping, shifting), namely G_1, G_2 and G_3 respectively. The classifications on those groups are performed by the classifier based on ResNet features and trained on the

Image group	Number of images	$A \pm I_a$	$L \pm I_l$
G_1	79,600	0.819 ± 0.003	0.187 ± 0.003
G_2	63,600	0.683 ± 0.004	0.319 ± 0.004
G_3	59,600	0.579 ± 0.004	0.423 ± 0.004

TABLE 4.13 – The classifications based on ResNet features performed on G_1, $G2$ and G_3 (images with naturalness decided by 7, 6 and 5 of 9 observers respectively).

high confidence data. The results of the classifications are presented in Table 4.13. It appears that the lower confidence data is, the lower classification performance is. As a matter of fact, the classification with G_1 is quite good (accuracy : 0.819, loss : 0.187) since the confidence data in G_1 is almost similar to the high confidence data of the training set. On G_2 and G_3, the opposite opinions of the observers about image naturalness decrease the confidence of the data and make the labels of the data unreliable, so the classifier does not work well in those cases.

4.5.2.10 INA with 3 classes : natural, unnatural and uncertain images

To go a step further, we tried to try to build a model classifying an image in three classes : natural, unnatural or uncertain based on the data obtained from the subjective experiment. Fig. 4.19 presents the structure of the model. It includes an input layer, an output layer and 7 hidden blocks and a feature extractor block. The output layer contains 3 neurons corresponding to the 3 classes (natural, uncertain and unnatural images). To train the model, the number of iterations is set to 150. The Adam optimizer is used and the loss function is the cross-entropy loss. The learning rate and the mini-batch size are set to 10^{-3}, 512 respectively. Regarding the feature extraction block, deep learned features seem quite good for natural / unnatural image classification and this classification is an extension of that one so transfered deep features are still chosen to perform this task. Different models including VGG16 [SZ15], Xception [Cho16], ResNet [He+15], NASNet large and NASNet mobile [Zop+17], MobileNet [How+17], Inception [Sze+15], DenseNet [HLW16], Inception ResNet [SIV16] pre-trained on the ImageNet dataset for the task of image classification without the prediction layers are considered as feature extractors. 3 image groups are extracted from the data. The first one contains images evaluated as natural by at least 8 of 9 observers (355 natural images). The second one contains images with naturalness labels decided by 5 of 9 observers (298 uncertain images). The last one contains images assessed as unnatural by at least 8 of 9 observers (335 unnatural images). The data augmentation presented in 4.3 is applied on the 3 groups to increase the number of samples. Those images are then split into a training set and a testing set (see details in Table 4.14). Those sets are then used to train and test the model with the different feature extractors. The results are presented in Table 4.15. It seems that deep features transferred from pre-trained models are not efficient in this case since the best accuracy belonging to features learned from the Dense Net model is only 0.658. The details of the classification based on DenseNet features is presented in Table 4.16. For natural images, 3,592

	Number of samples	
	Training set	Testing set
Natural images	58,600	12,400
Uncertain images	47,200	12,400
Unnatural images	58,600	12,400

TABLE 4.14 – Details of the training set and the testing set for natural / uncertain / unnatural image classification.

Natural / uncertain / unnatural image classification based on deep features learned from different deep models.		
Features learned from	A $\pm$ I_a	L $\pm$ I_l
NASNet Large model	0.499 $\pm$ 0.002	0.347 $\pm$ 0.002
Inception ResNet model	0.518 $\pm$ 0.002	0.351 $\pm$ 0.002
Inception model	0.520 $\pm$ 0.002	0.351 $\pm$ 0.002
Xception model	0.537 $\pm$ 0.002	0.332 $\pm$ 0.002
NASNet Mobile model	0.547 $\pm$ 0.002	0.332 $\pm$ 0.002
VGG16 model	0.589 $\pm$ 0.002	0.313 $\pm$ 0.002
MobileNet model	0.597 $\pm$ 0.002	0.298 $\pm$ 0.002
ResNet model	0.613 $\pm$ 0.002	0.264 $\pm$ 0.002
DenseNet model	0.658 $\pm$ 0.002	0.250 $\pm$ 0.002
Combination of all features	0.642 $\pm$ 0.002	0.243 $\pm$ 0.002

TABLE 4.15 – Natural / uncertain / unnatural image classification based on deep features learned from different deep models.

of 3,879 misclassified cases are uncertain images while for unnatural images, 1,170 of 1,985 wrong classifications are uncertain images. Additionally, the number of wrong classifications of uncertain images is also the highest (6,857 images against 3,879 and 1,985 images of natural and unnatural images respectively). Obviously, the wrong predictions are almost related to uncertain cases because the borders between the uncertain category and the others are not clear. Uncertain images locate in a middle range between natural images and unnatural images so they have some similar points to the both categories. The naturalness labels of those uncertain cases are mainly related to the observers' opinions.

4.6 Towards unnatural image understanding

According to the classification result analysis, it is confirmed that the feeling of unnaturalness comes from 2 main causes : strong unnaturalness clues detected by viewers' eyes mostly depicted with handcrafted features and image representation related to viewers' experience designed with learned features.

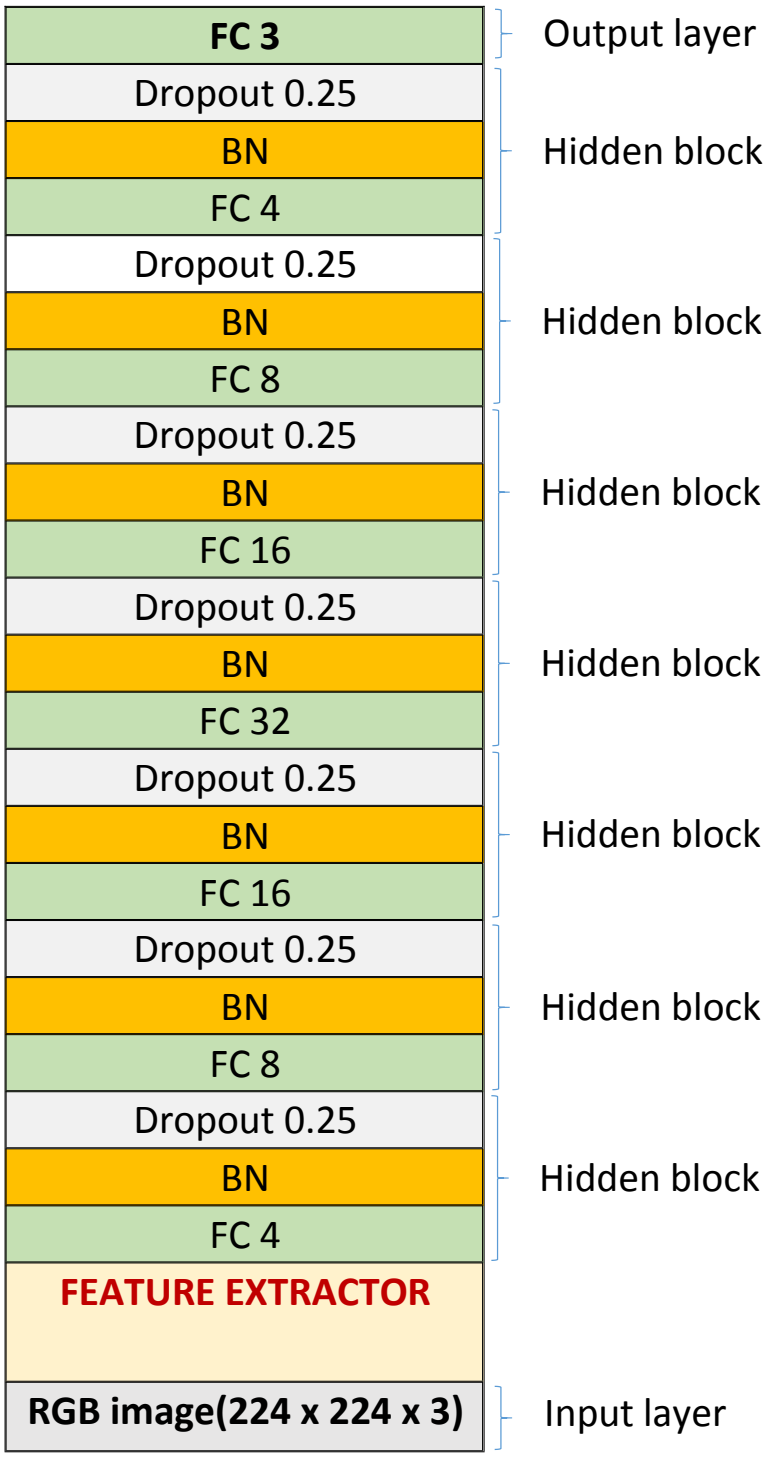


FIGURE 4.19 – Structure of the network designed for natural / uncertain /unnatural image classification. Features extracted from an RGB input image of size 224×224×3 by the feature extractor are passed through the layers to classify the image as natural, unnatural or uncertain. There are 7 hidden blocks with a fully connected layer, a batch normalization layer and a dropout layer in each block.

		Prediction		
		Natural	Uncertain	Unnatural
Ground truth	Natural	8521	3592	287
	Uncertain	3839	5543	3018
	Unnatural	815	1170	10,415
	A = 0.658 ± 0.002		L = 0.250 ± 0.002	

TABLE 4.16 – Natural / uncertain / unnatural image classification based on features learned from DenseNet model.

The first visible unnaturalness clue is color. It includes brightness, color saturation and hue. In general, it is impossible for a camera to cover the whole range of brightness of real scenes. By applying algorithms (TMOs, MEFs), the brightness range of a real scene is compressed and it leads to the fact that the brightness distribution, the brightness range and the brightness contrast of photos and those of real scenes are different. For instance, in the first row of Fig. 4.1, the left photo is too bright (over exposure) while the right one is too dark (under exposure). Additionally, the left image of the second row has a too high brightness contrast since some regions are too bright while some regions are too dark. Beside being compressed, the brightness also could be affected by using post processing algorithms. For example, when a photo has been taken under dark conditions, if the photographer wants to make it brighter, he / she might post-process the photo to increase the brightness. Generally, if the difference is insignificant, it might not be detected by viewers' eyes but if the difference is important, it becomes an artifact sign. TMOs, MEFs and post-processing re-produce brightness, color saturation and hue of images. An abnormal color saturation (too high or too low) could be detected by human eyes. Unusual hues in photos make photos unnatural to viewers. For instance, it is impossible to have orange sky as in Fig. 4.3 or dark blue sky as in Fig. 4.1.

Beside color, the second visible unnaturalness clue is the reproduction of details. In order to reproduce lost details, to enhance sharpness or to reduce noise in photos, some post-processing algorithms modify photo details. Those changes could lead to artifact signs such as blurriness, graininess, halo, dark band effects. Additionally, when combining multiple shoots taken under different exposures, MEFs try to preserve details coming from different images. Sometimes, the details are not combined well and some artifact clues such as motion blur, ghost effects are produced. TMOs generate unnatural details in a different way. When compressing color range, some TMOs try to preserve local contrast and global contrast in photos. The reduction of the dynamic range might produce artifact details (too sharp, contrast details, halo bands) or some details could be lost after mapping.

When the unnaturalness feeling comes from image viewer's representation, unnaturalness clues are not obvious. As a matter of fact, the observers compare scenes in photos to scenes retrieved from their memory (what they have seen) [WG14] to find differences and similarities, so assessment results depend on individual factors [GG12]. For example, some people think dark photos and bright photos are unnatural because they are not familiar with those scenes while some people disagree because they have seen similar scenes in few cloudy days or few

FIGURE 4.20 – Examples of uncertain images. The images in the left column are assessed as natural by 5 observers and as unnatural by 4 observers. The images in the right column are assessed as unnatural by 5 observers and as natural by 4 observers.

sunny days (see examples in Fig. 4.20). Another example is about the tree colors. Green colors of trees are not the same and they vary under different light conditions. However, some viewers fix a range of green colors for plants in their mind. Except those colors, they consider that other green colors are unnatural. In this case, when evaluating the naturalness of scenes with trees in photos, viewers often focus on 3 questions "What trees are they ?", "What are their colors ?" and "Do they and their colors match ?". As a result, it is not easy to design handcrafted features in this case because naturalness appears to be an individual feeling and in such a case deep learning helps us to learn the unnatural features common to every subject in the training set.

The 10 image groups of the dataset are merged into 5 categories (see Fig. 4.7) and each category is presented by a pair of values (X, Y) in which $X > Y$, X is the number of observers having the same opinion about the naturalness of an image while Y is the number of observers having the opposite opinion. According this definition, there are 5 categories including (9,0), (8,1), (7,2), (6,3) and (5,4). The naturalness label of each image is decided by the majority of the observers so it appears that the confidence of the labels in the category (9,0) is the highest and the confidence in the category (5,4) is the lowest. It appears that images with the highest confidence labels generally present obvious visible artifact whereas images with the lowest confidence labels are images on which the naturalness / unnaturalness feeling is more related to the viewer's experience.

Finally, the naturalness concept can definitely be defined based on 2 terms. The first one is memory color that reflects the typical color of an object that a beholder acquires through viewers' experience with that object [WG14]. The second term is obvious artifacts that can be recognized by eyes such as very high or very low contrast, too sharp details, loss details, artifact details (See Fig. 4.1). That is why the problem of naturalness assessment is so tricky.

4.7 Relations between image naturalness and image aesthetic

As presented in the introduction of the thesis, although there are relations between image aesthetic and image naturalness, they are 2 different concepts. Two experiments are organized to validate the relations between image naturalness and image aesthetic. The first one is to assess image aesthetic of the 355 natural images and 515 unnatural images (obtained from the subjective experiment presented in this chapter) by using the deep IAA model presented in chapter 3. The second one is to assess image naturalness of 10,524 high aesthetic images and 19,166 low aesthetic images (coming from the CUHKPQ dataset) by using the INA model presented in this chapter. The results of the 2 experiments are presented in Table 4.17. In the first experiment, it appears that only 17.7% of the natural images and 31.1% of the unnatural images are predicted as high aesthetic while a majority of them (82.3% of the natural images and 68.9% of the unnatural images) is assessed as low aesthetic. Considering image naturalness in the second experiment, 40.5% of the high aesthetic images are assessed as natural while the remaining high aesthetic images (59.5%) are assessed as unnatural. Considering low aesthetic images, 60.7% of them are evaluated as natural and the remains are indicated as unnatural. It

Image source	**Prediction**	
	High aesthetic	**Low aesthetic**
355 natural images	63 images (17.7%)	292 images (82.3%)
515 unnatural images	160 images (31.1%)	355 images (68.9%)
Image source	**Prediction**	
	Natural	**Unnatural**
10,524 high aesthetic images	4260 images (40.5%)	6264 images (59.5%)
19,166 low aesthetic images	11,631 images (60.7%)	7535 images (39.3%)

TABLE 4.17 – Predicted image naturalness for images coming from the image naturalness dataset and predicted image aesthetic for images coming from the CUHKPQ dataset.

proves that a high aesthetic image is not always natural and a natural image does not always mean a high aesthetic image because image naturalness and image aesthetic are not the same. Abusing enhancement methods that increase perceived aesthetic quality could provoke artifacts from imperceptible to obvious (over-enhancement) so the increase of image aesthetic could lead to the decrease of image naturalness (even decrease image quality generally). On the contrary, for example, comparing a photo re-produced by an adjustment method such as an TMO and other single exposure versions, the tone mapped photo could be more natural than other single exposure photos of the same scene with deep, lively and realistic colors and contrast. The adjusted photo could be more appealing and interesting because of the uniqueness (compared with normal single exposure images that cannot preserve high contrast, deep colors of the real scenes). In this case, naturalness is a factor that helps increasing image aesthetic quality. However, when a photo is too faithful and familiar to observers, there might be nothing interesting or special expected by viewers so it will not received high aesthetic evaluations from observers.

4.8 Conclusions

In this chapter, 2 main contributions have been presented. Firstly, an experiment of subjective image naturalness classification was organized. It was performed under strict experimental conditions. From 45 observers, over 17,000 subjective naturalness evaluations for 1900 SDR images have been obtained to establish a naturalness dataset for the purpose of analyzing photo naturalness automatically. Secondly, the image naturalness is evaluated in different ways using handcrafted features, features learned directly from CNN and transferred learned features. The experiments on the naturalness dataset point out the roles of the different feature types in the task of image naturalness evaluation. Handcrafted features are simple and quite efficient to detect obvious unnaturalness clues while deep learned features are complicated but get higher performance by detecting indescribable signs of unnaturalness. As a result, it has been demonstrated that handcrafted features and learned features are complementary. Finally, the relations between image aesthetic and image naturalness have been clarified. This might help to understand more the general concept of image quality (cf. Fig. 1.1).

Conclusions and perspectives

The main purposes of the conclusion chapter is to summarize the main contributions of this thesis and to present perspectives for future works.

In the thesis, we have proposed an extended definition of image quality and the two components of image quality including image aesthetic and image naturalness have been studied. On the one side, image aesthetic has been defined as the measure of how aesthetically a photo fulfills the observer's expectation. On the other side, image naturalness has been defined based on 2 aspects : viewers' color memory and perceptible artifacts. Two main problems related to image aesthetic and image naturalness have been tackled in this work. The first one is to study pre-processing operations for Image Aesthetic Assessment (IAA) : prior segmentation and prior image classification. Different approaches have been studied to perform those operations and the influences of the operations in IAA have been evaluated to propose an IAA model based on image classification and region segmentation. Secondly, image naturalness of tone-mapped images has been studied. A subjective experiment has been organized to collect subjective data about image naturalness. Then, different objective metrics have been validated on the collected subjective results to measure the efficiency of the metrics to find the best one for INA. Beside that, the relations between image naturalness and image aesthetic have been investigated and discussed.

Pre-processing methods for IAA : Two pre-processing methods have been studied in this study. The first one is region of interest extraction and the second one is large field / close-up image classification. They are the preparation steps before performing IAA.

About the first pre-processing method : Region Of Interest Extraction (ROIE), Regions Of Interest (ROIs) in this study are defined as regions attracting observers' eyes. ROIE has been studied with both handcrafted and deep learning based approaches. The experimental results prove that sharpness information only or color information only is not sufficient to precisely extract ROIs while the combination of them helps determining ROIs more efficiently. Both handcrafted and deep learning based approaches are effective to perform the task.

About the second pre-processing method : Large field / Close-up Image Classification (LCIC), large field images are considered as photos of a large field scene taken with a long distance from the camera to the scene while a close-up image is defined as a photo focusing on close-up objects captured with a short distance from the camera to the objects. Different types of features from simple to complex : EXIF features, handcrafted features and learned features have been considered to address the LCIC task. Those features have been investigated in terms of classification performance, feature complexity, computational cost. The experimental results prove that learned features are very efficient for this classification although they are complex, unintelligible and they require a strong GPU to reduce the computational time. An interesting point obtained from experimental results is about EXIF features. Although they are simple, EXIF features are efficient for the classification since it is possible to obtain a quite good classification accuracy by using 4 very simple EXIF features only.

IAA with or without prior image classification ? Considering IAA with prior LCIC, the key idea here is to assess image aesthetic of large field images and close-up images separately and 2 different aesthetic feature sets have been considered for the 2 image categories. IAA with prior LCIC has been compared with IAA without prior classification (aesthetic quality is assessed in the same way for any image). According to the experimental results, performing IAA for different image categories using different aesthetic features makes IAA more accurate than using the same aesthetic features for all images.

IAA with or without prior region segmentation ? Considering IAA with prior ROIE, based on the extracted ROIs and the extracted background from a given image, global features, ROI features and background features have been computed on the whole image, the ROIs and the background respectively. The influences of the 3 types of features and the combination of them in IAA have been evaluated in 2 cases : IAA for large field images only and IAA for close-up images only. The experimental results reflect that the combination of global features, ROI features and background features improves the performance of IAA for close-up images but the influence of ROI features and background features is insignificant in IAA for large field images. Therefore, performing prior ROIE before IAA or not depends on the particular situation.

IAA with prior image classification and region segmentation : According to the obtained results about the roles of LCIC and ROIE in IAA, a new IAA model based on LCIC and IAA has been introduced and evaluated. The experimental results have proved its efficiency compared to IAA without image classification and region segmentation.

Subjective experiment and objective metrics for image naturalness assessment : Image naturalness of tone-mapped images is a new topic since previous researches related to tone-mapped images mainly focus on general image quality. About this topic, the main purpose of this work is to assess image naturalness automatically but there was a big challenge because image naturalness datasets were not available. Therefore, there are 2 main contributions related to image naturalness. Firstly, in order to obtain subjective image naturalness data, an experiment of subjective image naturalness was organized. The obtained data has been used to establish an image naturalness dataset for the purpose of validating the performances of an automatic classification of image naturalness.

Secondly, different features of image naturalness classification have been introduced and validated on the collected subjective data. The experiments of those metrics performed on the naturalness dataset point out the roles of the different feature types in INA. According to the experimental results, handcrafted features are simple and quite efficient to the task while deep learned features are very complicated but get higher performances. Shallow learned features are not a good choice for analyzing image naturalness because of the modest performance among those objective metrics. The combination of handcrafted features and learned feature is a complement since the INA performance is improved. Finally, the relations between image aesthetic and image naturalness have been clarified. Based on the results predicted by the IAA model in Chapter 3 and the INA model in Chapter 4, it is clearly that although the 2 concepts influence to each other, image naturalness and image aesthetic are 2 different aspects defining image quality.

Efficiency of handcrafted features and learned features : In the studies of image aesthetic and image naturalness presented in the thesis, both handcrafted features and learned features have been exploited. The 2 types of features have been evaluated and compared to each other strictly in terms of complexity and performance. First of all, it is clear that learned features are very complex and unintelligible while handcrafted features are designed clearly and understandably. However, according to the results of the experiments, learned features are always more efficient than handcrafted features in IAA and INA. In the concern of computational cost, without GPU, the computational time for learned features is much higher than that of handcrafted features. However, with a good GPU, the computational time for learned features could be the same or even smaller than the computational time for handcrafted features. Both image aesthetic and image naturalness are abstract concepts depending on subjectivity (individual feelings) so it is not easy for handcrafted features to precisely describe or generalize those subjective opinions. On the other side, with features learned directly from images, they do not need to be defined explicitly so with those advantages, learned features are more suitable than handcrafted features in IAA and INA tasks.

Perspectives : According to the current results, there are 4 main directions for future researches. The first one is to develop algorithms able to modify image quality in order to improve it based on the 2 aspects : image aesthetic and image naturalness. Firstly, the global aesthetic features, ROI aesthetic features and background aesthetic features could be analyzed to find factors affecting image aesthetic significantly in order to enhance positive influence and reduce negative effects on image aesthetic. Similarly, naturalness features could be analyzed to find positive factors and negative factors affecting image naturalness. Detecting low aesthetic images, unnatural images could be considered as the first step to develop methods improving image quality by restoring naturalness of detected unnatural images and enhancing aesthetic quality of indicated low aesthetic images.

Secondly, the relations between image aesthetic, image naturalness and image quality could be digged more deeply. On the one side, aesthetic features could be analyzed to find the most relevant features towards image naturalness. The selected aesthetic features could be combined with naturalness features to see if the combination of them is efficient to the INA task or not. On the other side, naturalness features could be considered to select the most relevant features towards the IAA task. Similarly, the combination of the selected naturalness features and aesthetic features could be validated for the IAA task. Additionally, the most relevant features towards Image Quality Assessment (IQA) could be selected among the aesthetic features and the naturalness features to form 3 objective IQA metrics : one based on image aesthetic, another one based on image naturalness and the last one based on both image aesthetic and image naturalness. Those metrics could be validated on subjective IQA data to evaluate the roles of image aesthetic, image naturalness and the combination of them in IQA.

The next direction of future works could also focus more on regression problem by developing systems giving aesthetic scores and naturalness scores to images instead of a binary classification. The 2 systems could be exploited to develop an IQA system giving quality scores to images based on aesthetic scores and naturalness scores.

The last term is for image naturalness because it is a potential topic with many interesting

approaches. The next direction of the future work in this topic could be the organisation of an experiment with an HDR screen and HDR contents firstly to answer the question "are TMOs introducing unnaturalness in images or are the HDR images with unnatural artifacts even when displayed on an HDR screen ?" and secondly to analyze the similar points of naturalness features between HDR images and SDR images.

Bibliographie

[Ach+08] Radhakrishna Achanta et al. “Salient Region Detection and Segmentation”. In : *Computer Vision Systems* 5008.2008 (2008), p. 66-75 (cf. p. 39).

[Ach+12] Radhakrishna Achanta et al. “SLIC superpixels compared to state-of-the-art superpixel methods”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), p. 2274-2281 (cf. p. 27).

[ADG12] Pinaki Pratim Acharjya, Ritaban Das et Dibyendu Ghoshal. “A study on image edge detection using the gradients”. In : *International Journal of Scientific and Research Publications* 2.12 (2012), p. 1-5 (cf. p. 105).

[ASG15] Tunc Ozan Aydin, Aljoscha Smolic et Markus Gross. “Automated aesthetic analysis of photographic images”. In : *IEEE Transactions on Visualization and Computer Graphics* 21.1 (2015), p. 31-42 (cf. p. 27, 29, 30, 39, 42, 53, 68-70).

[Ash02] Michael Ashikhmin. “A Tone Mapping Algorithm for High Contrast Images”. In : *Proceedings of the 13th Eurographics Workshop on Rendering*. EGRW ’02. Pisa, Italy : Eurographics Association, 2002, p. 145-156 (cf. p. 90, 94, 98, 99).

[Bas+19] T. Bashford-Rogers et al. “Learning Preferential Perceptual Exposure for HDR Displays”. In : *IEEE Access* 7 (2019), p. 36800-36809 (cf. p. 87).

[BL04] Matthew Boutell et Jiebo Luo. “Photo classification by integrating image content and camera metadata”. In : *Proceedings - International Conference on Pattern Recognition* 4.January (2004), p. 901-904 (cf. p. 48).

[BL05] Matthew Boutell et Jiebo Luo. “Beyond pixels : Exploiting camera metadata for photo classification”. In : *Pattern Recognition* 38.6 (2005). Image Understanding for Photographs, p. 935 -946 (cf. p. 48).

[Bru+13] Kjell Brunnström et al. *Qualinet White Paper on Definitions of Quality of Experience*. Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013. Mar. 2013 (cf. p. 3, 15).

[BZM07] A. Bosch, A. Zisserman et X. Munoz. “Image Classification using Random Forests and Ferns”. In : *2007 IEEE 11th International Conference on Computer Vision*. 2007, p. 1-8 (cf. p. 48).

[Čad+08] Martin Čadík et al. “Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes”. In : *Computers & Graphics* 32 (3 2008), p. 330-349 (cf. p. 94).

[Che+13] Ming Ming Cheng et al. “Efficient salient region detection with soft image abstraction”. In : 2013, p. 1529-1536 (cf. p. 27).

[Che+15] Ming Ming Cheng et al. “Global contrast based salient region detection”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015), p. 569-582 (cf. p. 28).

[Cho+09] Seo Young Choi et al. "Investigation of large display color image appearance–III : Modeling image naturalness". In : *Journal of Imaging Science and Technology* 53.3 (2009), p. 31104-1 (cf. p. 91, 92).

[Cho16] François Chollet. "Xception : Deep Learning with Depthwise Separable Convolutions". In : *Computing Research Repository* abs/1610.02357 (2016) (cf. p. 108, 124).

[Cor+16] Marcella Cornia et al. "A deep multi-level network for saliency prediction". In : *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, p. 3488-3493 (cf. p. 28).

[Cre+07] Frederique Crete et al. "The blur effect : perception and estimation with a new no-reference perceptual blur metric". In : *Vol. 6492, Human Vision and Electronic Imaging XII* 6492.May 2016 (2007), p. 64920I-64920I-11 (cf. p. 29).

[CS05] Martin Cadik et Pavel Slavik. "The Naturalness of Reproduced High Dynamic Range Images". In : *Proceedings of the Ninth International Conference on Information Visualisation*. IV '05. Washington, DC, USA : IEEE Computer Society, 2005, p. 920-925 (cf. p. 91, 92, 101).

[Dat+06] Ritendra Datta et al. "Studying Aesthetics in Photographic Images Using a Computational Approach". In : t. 3. Mai 2006, p. 288-301 (cf. p. 53, 67, 70).

[DD02] Frédo Durand et Julie Dorsey. "Fast bilateral filtering for the display of high-dynamic-range images". In : *ACM transactions on graphics (TOG)*. T. 21. 3. ACM. 2002, p. 257-266 (cf. p. 90, 94, 98, 99).

[DE96] Thomas J DiCiccio et Bradley Efron. "Bootstrap confidence intervals". In : *Statistical science* (1996), p. 189-212 (cf. p. 39, 111).

[DLT17] Yubin Deng, Chen Change Loy et Xiaoou Tang. "Image aesthetic assessment : An experimental survey". In : *IEEE Signal Processing Magazine* 34.4 (2017), p. 80-106 (cf. p. 67, 90).

[DM97] Paul E. Debevec et Jitendra Malik. "Recovering High Dynamic Range Radiance Maps from Photographs". In : *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. New York, NY, USA : ACM Press/Addison-Wesley Publishing Co., 1997, p. 369-378 (cf. p. 87, 94).

[DOB11] Sagnik Dhar, Vicente Ordonez et Tamara L Berg. "High level describable attributes for predicting aesthetics and interestingness". In : *CVPR 2011*. IEEE. 2011, p. 1657-1664 (cf. p. 67).

[Fai07] Mark D. Fairchild. "The HDR Photographic Survey". In : *Color Imaging Conference*. Jan. 2007, p. 233-238 (cf. p. 94).

[FLW02] Raanan Fattal, Dani Lischinski et Michael Werman. "Gradient domain high dynamic range compression". In : *ACM transactions on graphics (TOG)*. T. 21. 3. ACM. 2002, p. 249-256 (cf. p. 90, 94, 98, 99).

[Fu+13a] Keren Fu et al. “Geodesic saliency propagation for image salient region detection”. In : *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings* (2013), p. 3278-3282 (cf. p. 31).

[Fu+13b] Keren Fu et al. “Superpixel based color contrast and color distribution driven salient object detection”. In : *Signal Processing : Image Communication* 28.10 (2013), p. 1448-1463 (cf. p. 28).

[Gao+10] Xinbo Gao et al. “Image quality assessment and human visual system”. In : *Visual Communications and Image Processing 2010*. T. 7744. International Society for Optics et Photonics. 2010, 77440Z (cf. p. 1, 13).

[GG12] Jeroen Granzier et Karl Gegenfurtner. “Effects of Memory Colour on Colour Constancy for Unknown Coloured Objects”. In : *i-Perception* 3 (avr. 2012), p. 190-215 (cf. p. 127).

[Gos05] A Ardeshir Goshtasby. “Fusion of multi-exposure images”. In : *Image and Vision Computing* 23.6 (2005), p. 611-618 (cf. p. 87).

[Gu+16] Ke Gu et al. “Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure”. In : *IEEE Transactions on Multimedia* 18.3 (2016), p. 432-443 (cf. p. 91, 92).

[Guo+17] T. Guo et al. “Simple convolutional neural network on image classification”. In : *2017 IEEE 2nd International Conference on Big Data Analysis*. 2017, p. 721-724 (cf. p. 48).

[HCC08] H. Huang, Y. Chen et S. Chen. “Copyright Protection for Images with EXIF Metadata”. In : *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2008, p. 239-242 (cf. p. 48).

[He+15] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In : *Computing Research Repository* abs/1512.03385 (2015) (cf. p. 108, 114, 124).

[He+18] Shiyi He et al. “Reinforced Multi-Label Image Classification by Exploring Curriculum”. In : *The Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cf. p. 48).

[HLW16] Gao Huang, Zhuang Liu et Kilian Q. Weinberger. “Densely Connected Convolutional Networks”. In : *Computing Research Repository* abs/1608.06993 (2016) (cf. p. 108, 124).

[How+17] Andrew G. Howard et al. “MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications”. In : *Computing Research Repository* abs/1704.04861 (2017) (cf. p. 108, 124).

[HS11] Peter D. Hiscocks et P. Eng Syscomp. *Measuring Luminance with a Digital Camera*. 2011 (cf. p. 51).

[Jia+11] Huaizu Jiang et al. “Automatic Salient Object Segmentation Based on Context and Shape Prior”. In : 2011, p. 1-12 (cf. p. 27).

[Jia+18] G. Jiang et al. “Blind Tone-Mapped Image Quality Assessment Based on Brightest/Darkest Regions, Naturalness and Aesthetics”. In : *IEEE Access* 6 (2018), p. 2231-2240 (cf. p. 91, 92).

[JSS16] Bin Jin, Maria V Ortiz Segovia et Sabine Susstrunk. “Image aesthetic predictors based on weighted CNNs”. In : *Proceedings - International Conference on Image Processing, ICIP*. T. 2016-Augus. 2016, p. 2291-2295 (cf. p. 56).

[KBK11] K. Kim, J. Bae et J. Kim. “Natural hdr image tone mapping based on retinex”. In : *IEEE Transactions on Consumer Electronics* 57.4 (2011), p. 1807-1814 (cf. p. 87, 90).

[Kha+18] I. R. Khan et al. “A Tone-Mapping Technique Based on Histogram Using a Sensitivity Model of the Human Visual System”. In : *IEEE Transactions on Industrial Electronics* 65.4 (2018), p. 3469-3479 (cf. p. 87, 94, 98, 99).

[Kon+16] Shu Kong et al. “Photo aesthetics ranking network with attributes and content adaptation”. In : *European Conference on Computer Vision*. Springer. 2016, p. 662-679 (cf. p. 68).

[Kor+14] Pavel Korshunov et al. “Crowdsourcing-based Evaluation of Privacy in HDR Images”. In : *Optics, Photonics, And Digital Technologies For Multimedia Applications Iii*. Proceedings of SPIE 9138 (2014), p. 11 (cf. p. 94).

[KR92] Kenji Kira et Larry A. Rendell. “A Practical Approach to Feature Selection”. In : *Proceedings of the Ninth International Workshop on Machine Learning*. ML92. Aberdeen, Scotland, United Kingdom : Morgan Kaufmann Publishers Inc., 1992, p. 249-256 (cf. p. 53).

[Kra+17] L. Krasula et al. “Preference of Experience in Image Tone-Mapping : Dataset and Framework for Objective Measures Comparison”. In : *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2017), p. 64-74 (cf. p. 1, 13, 94).

[KSH17] Alex Krizhevsky, Ilya Sutskever et Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In : *Commun. ACM* 60.6 (mai 2017), p. 84-90 (cf. p. 56, 105).

[KTJ06] Yan Ke, Xiaoou Tang et Feng Jing. “The design of high-level features for photo quality assessment”. In : t. 1. 2006, p. 419-426 (cf. p. 53, 70).

[Kun+17] D. Kundu et al. “Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures”. In : *IEEE Transactions on Image Processing* 26.10 (2017), p. 4725-4740 (cf. p. 94).

[LE+19] Quyet Tien LE et al. “Large Field/Close-Up Image Classification : From Simple to Very Complex Features”. In : *Computer Analysis of Images and Patterns*. Sous la dir. de Mario Vento et Gennaro Percannella. T. 11679. Lecture Notes in Computer Science. Proceedings of the 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Part II. Springer, 2019, 532-543, 10.1007/978-3-030-29891-3_47 . hal-02368500 (cf. p. 27, 54).

[LE+20] Quyet Tien LE et al. “Study of naturalness in tone-mapped images”. In : *Computer Vision and Image Understanding* 196 (2020), 102971. 10.1016/j.cviu.2020.102971. hal-02568771 (cf. p. 3, 15, 90).

[LF10] Zhong Li et Jianping Fan. “Exploit Camera Metadata for Enhancing Interesting Region Detection and Photo Retrieval”. In : *Multimedia Tools Appl.* 46.2-3 (jan. 2010), p. 207-233 (cf. p. 48).

[LH16] Nian Liu et Junwei Han. “Dhsnet : Deep hierarchical saliency network for salient object detection”. In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, p. 678-686 (cf. p. 28).

[Liu+17] Zexia Liu et al. “Background priors based saliency object detection”. In : 2017 (cf. p. 31).

[LRP97] Gregory Ward Larson, Holly Rushmeier et Christine Piatko. “A visibility matching tone reproduction operator for high dynamic range scenes”. In : *IEEE Transactions on Visualization and Computer Graphics* 3.4 (1997), p. 291-306 (cf. p. 90, 94, 98, 99).

[LT08] Yiwen Luo et Xiaoou Tang. “Photo and Video Quality Evaluation : Focusing on the subject”. In : t. 8. 08. 2008, p. 386-399 (cf. p. 27, 53, 70).

[LT16] Hao Lv et Xinmei Tian. “Learning relative aesthetic quality with a pairwise approach”. In : *International Conference on Multimedia Modeling*. Springer. 2016, p. 493-504 (cf. p. 69).

[Lu+15] Xin Lu et al. “Rating image aesthetics using deep learning”. In : *IEEE Transactions on Multimedia* 17.11 (2015), p. 2021-2034 (cf. p. 68).

[LW07] D. Lu et Q. Weng. “A survey of image classification methods and techniques for improving classification performance”. In : *International Journal of Remote Sensing* 28.5 (2007), p. 823-870 (cf. p. 48).

[LY15] Guanbin Li et Yizhou Yu. “Visual saliency based on multiscale deep features”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, p. 5455-5463 (cf. p. 28).

[Mar10] David Marr. *Vision : A computational investigation into the human representation and processing of visual information*. MIT press, 2010 (cf. p. 1, 13).

[Mar+11] Luca Marchesotti et al. “Assessing the aesthetic quality of photographs using generic image descriptors”. In : *2011 international conference on computer vision*. IEEE. 2011, p. 1784-1791 (cf. p. 68).

[MEMS14] Pedram Mohammadi, Abbas Ebrahimi-Moghadam et Shahram Shirani. “Subjective and Objective Quality Assessment of Image : A Survey”. In : *CoRR* abs/1406.7799 (2014) (cf. p. 90).

[Mit97] Thomas M. Mitchell. *Machine Learning*. 1re éd. New York, NY, USA : McGraw-Hill, Inc., 1997 (cf. p. 39, 111).

[MJL16] Long Mai, Hailin Jin et Feng Liu. “Composition-preserving deep photo aesthetics assessment”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 497-506 (cf. p. 68).

[MKVR09] Tom Mertens, Jan Kautz et Frank Van Reeth. “Exposure fusion : A simple and practical alternative to high dynamic range photography”. In : *Computer graphics forum*. T. 28. 1. Wiley Online Library. 2009, p. 161-171 (cf. p. 87).

[MM15] Eftichia Mavridaki et Vasileios Mezaris. "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography". In : *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, p. 887-891 (cf. p. 68).

[MMP12] N. Murray, L. Marchesotti et F. Perronnin. "AVA : A large-scale database for aesthetic visual analysis". In : *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, p. 2408-2415 (cf. p. 69).

[Nar+13] Manish Narwaria et al. "Tone mapping-based high-dynamic-range image compression : study of optimization criterion and perceptual quality". In : *Optical Engineering* 52.10 (2013) (cf. p. 94).

[Nis+11] Masashi Nishiyama et al. "Aesthetic quality classification of photographs based on color harmony". In : *CVPR 2011*. IEEE. 2011, p. 33-40 (cf. p. 68).

[Ots79] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms". In : *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), p. 62-66 (cf. p. 31).

[Per+12] Federico Perazzi et al. "Saliency filters : Contrast based filtering for salient region detection". In : 2012, p. 733-740 (cf. p. 27, 31, 39, 42).

[PK10] Fabrizio Pece et Jan Kautz. "Bitmap movement detection : HDR for dynamic scenes". In : *2010 Conference on Visual Media Production*. IEEE. 2010, p. 1-8 (cf. p. 94, 98, 99).

[PSA16] Sujoy Paul, Ioana S Sevcenco et Panajotis Agathoklis. "Multi-exposure and multi-focus image fusion in gradient domain". In : *Journal of Circuits, Systems and Computers* 25.10 (2016), p. 1650123 (cf. p. 94, 98, 99).

[PSh15] Federico Perazzi et Olga Sorkine-hornung. "Efficient Salient Foreground Detection for Images and Video using Fiedler Vectors". In : 2015 (cf. p. 31).

[PY10] S. J. Pan et Q. Yang. "A Survey on Transfer Learning". In : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), p. 1345-1359 (cf. p. 56, 105, 108).

[RBF95] Huib de Ridder, Frans JJ Blommaert et Elena A Fedorovskaya. "Naturalness and image quality : chroma and hue variation in color images of natural scenes". In : *Human Vision, Visual Processing, and Digital Display VI*. T. 2411. International Society for Optics et Photonics. 1995, p. 51-62 (cf. p. 91, 92).

[RC09] Shanmuganathan Raman et Subhasis Chaudhuri. "Bilateral Filter Based Compositing for Variable Exposure Photography." In : *Eurographics (short papers)*. 2009, p. 1-4 (cf. p. 94, 98, 99).

[Rei+02] Erik Reinhard et al. "Photographic Tone Reproduction for Digital Images". In : *ACM Trans. Graph.* 21.3 (juil. 2002), p. 267-276 (cf. p. 90, 94, 98, 99).

[Rei+10] Erik Reinhard et al. *High dynamic range imaging : acquisition, display, and image-based lighting, ISBN : 9780123749147, 9780080957111*. 2e éd. Morgan Kaufmann publisher, 2010 (cf. p. 87).

[Rid96] Huib de Ridder. "Naturalness and image quality : saturation and lightness variation in color images of natural scenes". In : *Journal of imaging science and technology* 40.6 (1996), p. 487-493 (cf. p. 91, 92).

[RW17] Waseem Rawat et Zenghui Wang. "Deep Convolutional Neural Networks for Image Classification : A Comprehensive Review". In : *Neural Computation* 29.9 (2017), p. 2352-2449 (cf. p. 48).

[See+04] Helge Seetzen et al. "High Dynamic Range Display Systems". In : *ACM Transactions on Graphics* 23 (mai 2004) (cf. p. 87).

[SIV16] Christian Szegedy, Sergey Ioffe et Vincent Vanhoucke. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In : *Computing Research Repository* abs/1602.07261 (2016) (cf. p. 108, 124).

[SS16] S. Suran et Sreekumar K. "Automatic aesthetic quality assessment of photographic images using deep convolutional neural network". In : *2016 International Conference on Information Science (ICIS)*. 2016, p. 77-82 (cf. p. 69).

[SZ15] Karen Simonyan et Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In : *Computing Research Repository* abs/1409.1556 (2015) (cf. p. 56, 108, 124).

[Sze+15] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In : *Computing Research Repository* abs/1512.00567 (2015) (cf. p. 108, 124).

[Tec02] Technical Standardization Committee on AV & IT Storage Systems and Equipment. *Exchangeable image file format for digital still cameras : Exif Version 2.2*. Rapp. tech. JEITA CP-3451. 2002 (cf. p. 48).

[Tia+15] Xinmei Tian et al. "Query-dependent aesthetic model with deep learning for photo quality assessment". In : *IEEE Transactions on Multimedia* 17.11 (2015), p. 2035-2048 (cf. p. 68, 69).

[TLW13] Xiaoou Tang, Wei Luo et Xiaogang Wang. "Content-based photo quality assessment". In : *IEEE Transactions on Multimedia* 15.8 (2013), p. 1930-1943 (cf. p. 27, 36, 39, 42, 56, 74, 77).

[TM18] Hossein Talebi et Peyman Milanfar. "NIMA : Neural Image Assessment". In : *IEEE Transactions on Image Processing* 27 (2018), p. 3998-4011 (cf. p. 56).

[TM98] C. Tomasi et R. Manduchi. "Bilateral Filtering for Gray and Color Images". In : *International Conference on Computer Vision* (1998), p. 839-846 (cf. p. 30).

[Ton+15] Na Tong et al. "Salient Object Detection via Bootstrap Learning Supplementary Materials". In : *IEEE Conference on Computer Vision and Pattern Recognition* (2015), p. 2012 (cf. p. 28).

[Ton+16] S. Tong et al. "Visual attention inspired distant view and close-up view classification". In : *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, p. 2787-2791 (cf. p. 48).

[Uni08] International Telecommunication Union. *ITU-T Recommendation E.800. Definitions of terms related to quality of service*. 2008, p. 15 (cf. p. 15).

[Vai+99] a. Vailaya et al. "Content-based hierarchical classification of vacation images". In : t. 1. 2. 1999, p. 518-523 (cf. p. 53, 70).

[Wan] "Hierarchical Image Classification Using Support Vector Machines". In : *Asian Conference on Computer Vision* January (2002), p. 23-25 (cf. p. 48, 60).

[Wan+16a] Weining Wang et al. "A multi-scene deep learning model for image aesthetic evaluation". In : *Signal Processing : Image Communication* 47 (2016), p. 511-518 (cf. p. 69).

[Wan+16b] Zhangyang Wang et al. "Brain-inspired deep networks for image aesthetics assessment". In : *arXiv preprint arXiv :1601.04155* (2016) (cf. p. 69).

[WG14] Christoph Witzel et Karl Gegenfurtner. "Memory Color". In : *Encyclopedia of Color Science and Technology*. Sous la dir. de Ronnier Luo. New York, NY : Springer New York, 2014, p. 1-7 (cf. p. 127, 129).

[WL09] Lai-Kuan Wong et Kok-Lim Low. "Saliency-enhanced image aesthetics class prediction". In : *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, p. 997-1000 (cf. p. 67, 69).

[WT97] Brian Wandell et Stephen Thomas. "Foundations of vision". In : *Psyccritiques* 42.7 (1997) (cf. p. 1, 13).

[YMB17] Charles Yaacoub, Jad Melhem et Petra Bilane. "A No-Reference Metric for Quality Assessment of Tone-Mapped High Dynamic Range Images". In : *International Journal of Applied Engineering Research* 12 (juin 2017), p. 2598-2603 (cf. p. 91, 92).

[YW13] H. Yeganeh et Z. Wang. "Objective Quality Assessment of Tone-Mapped Images". In : *IEEE Transactions on Image Processing* 22.2 (2013), p. 657-667 (cf. p. 94).

[Zha+15] Rui Zhao et al. "Saliency detection by multi-context deep learning". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, p. 1265-1274 (cf. p. 28).

[Zhu+14] C. Zhuang et al. "Anaba : An obscure sightseeing spots discovering system". In : *2014 IEEE International Conference on Multimedia and Expo (ICME)* (2014), p. 1-6 (cf. p. 48, 60).

[Zop+17] Barret Zoph et al. "Learning Transferable Architectures for Scalable Image Recognition". In : *Computing Research Repository* abs/1707.07012 (2017) (cf. p. 108, 124).

[ZZC13] Zhenzhu Zheng, Yun Zhang et Qian Chen. "Salient Object Detection Using Background and Foreground Prior". In : 2013 (cf. p. 28, 31, 39, 42).

Résumé — Dans cette thèse, les principales contributions consistent à étudier 2 aspects principaux de la qualité d'image, l'esthétique de l'image, le naturel de l'image ainsi que les relations entre ces 2 concepts. Plus précisément, l'esthétique de l'image est la mesure de la façon dont une photo répond esthétiquement aux attentes de l'observateur, tandis que la définition du naturel de l'image est à la fois liée aux artefacts induits par certains algorithmes de traitement d'image et en conséquence au sentiment individuel dont une image correspond à la mémoire qu'on en a.

Afin d'élaborer un modèle d'évaluation de l'esthétique d'une image, nous abordons dans cette thèse l'impact de deux prétraitements possibles à savoir la classification (champs large/champs proche) mais aussi la segmentation (extraction des régions d'intérêt), puis nous les comparons aux modèles sans étapes de prétraitement. Dans le même temps, différents modèles basés soit sur des caractéristiques extraites des images, soit sur des caractéristiques apprises ont été étudiés aux fins d'estimation esthétique de l'image. Sur la base des différents résultats obtenus, un modèle d'évaluation esthétique d'image basé sur la classification d'image et la segmentation de région a été introduit et évalué.

Dans le cadre de l'étude du naturel de l'image, notre travail a porté sur les effets apportés par les algorithmes de mapping de tonalité permettant de passer des images HDR aux images SDR affichées sur les écrans standards. Nous nous sommes intéressés aux méthodologies à la fois subjectives et objectives. Une expérience subjective a été organisée pour recueillir des évaluations humaines sur le naturel de l'image. Ensuite, divers algorithmes objectifs ont été validés sur les données subjectives collectées pour la tâche d'évaluation du naturel de l'image. Ce travail se concentre sur le problème du développement d'un modèle dans un premier temps pour estimer si une image semble naturelle ou non aux humains puis le second objectif est d'essayer de comprendre comment le sentiment de "non-naturel" est induit par une photo : "Y a-t-il des indices spécifiques au non-naturel ou est-ce un sentiment général en regardant une photo ?". Enfin, les relations entre les 2 aspects : esthétique de l'image et naturel de l'image ont été évaluées et discutées.

Mots clés : qualité d'image, esthétique de l'image, naturel de l'image, région d'intérêt, classification d'image à champs large et champs proche, image HDR, image SDR, opérateur de mappage de ton, caractéristiques images, caractéristiques apprises, apprentissage par transfert, EXIF, réseau neuronal convolutif, expérience subjective, métrique objective.

Abstract — In this thesis, the main contributions are to study 2 main aspects of image quality including image aesthetic, image naturalness and the relations between the 2 concepts. More specifically, image aesthetic is the measure of how aesthetically a photo fulfills the observer's expectation while the image naturalness definition is both related to artifacts

induced by some image processing algorithms and to the individual feeling about how a picture matches with image memory. On the side of image aesthetic, the thesis deals with the problem of evaluating the roles of pre-processing operations in image aesthetic assessment. Image aesthetic assessment models based on prior image segmentation (region of interest extraction) and prior image classification (large field / close-up image classification) have been developed and compared with image aesthetic assessment models without pre-processing stages. At the same time different models base either on handcrafted features or learned features have been studied for the purpose of image aesthetic estimation. Based on the obtained results, an image aesthetic assessment model based on image classification and region segmentation has been introduced and evaluated. On the side of image naturalness, image naturalness of standard dynamic range images, especially tone-mapped images have been studied with both subjective and objective methodologies. A subjective experiment has been organized to collect human evaluations about image naturalness first. Then, various objective algorithms have been validated on the collected subjective data for the image naturalness assessment task. This work focuses on the problem of developing a model firstly to estimate if an image looks natural or not to humans and the second purpose is to try to understand how the unnaturalness feeling is induced by a photo : "Are there specific unnaturalness clues or is unnaturalness a general feeling when looking at a photo ?". Finally, the relations between the 2 aspects : image aesthetic and image naturalness have been evaluated and discussed.

Keywords : image quality, image aesthetic, image naturalness, region of interest, large field close-up image classification, HDR image, SDR image, tone mapping operator, handcrafted features, learned features, transfer learning, EXIF, convolutional neural network, subjective experiment, objective metric.

GIPSA-lab, 11 rue des Mathématiques, Grenoble Campus BP 46
Grenoble