Perceptual object of interest recognition : application to the interpretation of instrumental activities of daily living for dementia studies

Vincent Buso

université de BORDEAUX

THÈSE

PRÉSENTÉE

L'UNIVERSITÉ BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Vincent Buso**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Reconnaissance Perceptuelle Des Objets D'Intêret: Application A L'Interprétation Des Activités Instrumentales De La Vie Quotidienne Pour Les Etudes De Démence

---

Soutenue le : 30 11 2015

Après avis des rapporteurs :

M. Matthieu J. CORD	Professeur, UPMC-Université de Sorbonne, Paris
Mme Christine FERNANDEZ-MALOIGNE	Professeur, XLIM-SIC, Poitiers

Devant la commission d'examen composée de :

Mme Jenny BENOIS-PINEAU	Professeur, Université Bordeaux	Directrice de thèse
M. Matthieu J. CORD	Professeur, UPMC-Université de Sorbonne, Paris	Rapporteur
M. Jean-Philippe DOMENGER	Professeur, Université Bordeaux	Président
M. Michael DORR	Research group leader, TUEIMMK, Munich	Examinateur
Mme Christine FERNANDEZ-MALOIGNE	Professeur, XLIM-SIC, Poitiers	Rapporteur
M. Ivan LAPTEV	Directeur de recherche, INRIA, Paris	Examinateur

– 2015 –

Acknowledgements

First I would like to thank my supervisor Jenny Benois-Pineau for allowing me to work on such an interesting subject but mostly for guiding me throughout these three years with her scientific rigor and numerous advice. I specifically appreciate the liberty she left me to explore different tracks and her high availability during these three years.

I am truly thankful that Prof. Matthieu Cord and Prof. Christine Fernandez-Maloigne have accepted to review this manuscript. I also would like to thank the members of the jury Prof. Jean-Philippe Domenger, Dr. Michael Dorr and Dr. Ivan Laptev to be part of my thesis defense committee.

I would also like to thank all the fascinating people I met and came to work with in the context of the Dem@care european project. In particular I am grateful for having closely worked with the members IMS laboratory: Prof. Yannick Berthoumieu, MC. Remi Megret, Gaelle, Etienne... Last but not least from this laboratory is my fellow Dr. Guillaume Bourmaud and his girlfriend Dr. Cornelia Vacar for whom this sentence is definitely too small to express all my gratitude.

Regarding my own laboratory, the LaBRI, I will never forget all the people that I met and who all left their mark somehow in this work. I would like to thank in particular Dr. Hugo Boujut who was there to take good care of the newcomer that I was and taught me a great deal in coding. Another person I am truly thankful for having crossed my path is Dr. Ivan Gonzalez-Diaz whose mind, ideas, pedagogy and modesty are yet to be matched in my opinion. Of course I will never forget the time spent in the company of Mr. Aurelien Gibiat, with whom I go way back and share so much similar interests. And of course I will never forget you: Olfa, Souad, Pierre-Marie, Jerome, Alix, Antoine... again thanks to you all. I do not want to forget all the people from the administration and the system team who helped me so many times and definitely contribute to the good mood at the LaBRI, thank you all.

I deeply thank you, Valeria, for the time, support and love you always managed to give me, even during the hard final times of the writing.

And finally I cannot finish without expressing all the gratitude I have for my family for their support during these three years but really, simply for all these years of my life, period.

Reconnaissance Perceptuelle Des Objets D'Intêret: Application A L'Interprétation Des Activités Instrumentales De La Vie Quotidienne Pour Les Etudes De Démence

Résumé: Cette thèse est motivée par le diagnostic, l'évaluation, la maintenance et la promotion de l'indépendance des personnes souffrant de maladies démentielles pour leurs activités de la vie quotidienne. Dans ce contexte nous nous intéressons à la reconnaissance automatique des activités de la vie quotidienne.

L'analyse des vidéos de type égocentriques (où la caméra est posée sur une personne) a récemment gagné beaucoup d'intérêt en faveur de cette tâche. En effet de récentes études démontrent l'importance cruciale de la reconnaissance des objets actifs (manipulés ou observés par le patient) pour la reconnaissance d'activités et les vidéos égocentriques présentent l'avantage d'avoir une forte différenciation entre les objets actifs et passifs (associés à l'arrière plan). Une des approches récentes envers la reconnaissance des éléments actifs dans une scène est l'incorporation de la saillance visuelle dans les algorithmes de reconnaissance d'objets. Modéliser le processus sélectif du système visuel humain représente un moyen efficace de focaliser l'analyse d'une scène vers les endroits considérés *d'intérêts* ou *saillants*, qui, dans les vidéos égocentriques, correspondent fortement aux emplacements des objets d'intérêt. L'objectif de cette thèse est de permettre au systèmes de reconnaissance d'objets de fournir une détection plus précise des objets d'intérêts grâce à la saillance visuelle afin d'améliorer les performances de reconnaissances d'activités de la vie de tous les jours. Cette thèse est menée dans le cadre du projet Européen Dem@care.

Concernant le vaste domaine de la modélisation de la saillance visuelle, nous étudions et proposons une contribution à la fois dans le domaine "Bottom-up" (regard attiré par des stimuli) que dans le domaine "Top-down" (regard attiré par la sémantique) qui ont pour but d'améliorer la reconnaissance d'objets actifs dans les vidéos égocentriques. Notre première contribution pour les modèles Bottom-up prend racine du fait que les observateurs d'une vidéo sont normalement attirés par le centre de celle-ci. Ce phénomène biologique s'appelle le *biais central*. Dans les vidéos égocentriques cependant, cette hypothèse n'est plus valable. Nous proposons et étudions des modèles de saillance basés sur ce phénomène de biais non-central. Les modèles proposés sont entrainés à partir de fixations d'œil enregistrées et incorporées dans des modèles spatio-temporels. Lorsque comparés à l'état-de-l'art des modèles Bottom-up, ceux que nous présentons montrent des résultats prometteurs qui illustrent la nécessité d'un modèle géométrique biaisé non-centré dans ce type de vidéos. Pour notre contribution dans le domaine Top-down, nous présentons un modèle probabiliste d'attention visuelle pour la reconnaissance d'objets manipulés dans les vidéos égocentriques. Bien que les bras soient souvent source d'occlusion des objets et considérés comme un fardeau, ils deviennent un atout dans notre approche. En effet nous extrayons à la fois des caractéristiques globales et locales permettant d'estimer leur disposition géométrique. Nous intégrons cette information dans un modèle probabiliste, avec équations de mise a jour pour optimiser la vraisemblance du modèle en fonction de ses paramètres et enfin générons les cartes

d'attention visuelle pour la reconnaissance d'objets manipulés. Ce modèle probabiliste Top-down permet de définir de nouveaux résultats d'état-de-l'art pour la reconnaissance d'objets manipulés pour ce genre de vidéos.

En ce qui concerne le fait de reconnaître les objets actifs, nous proposons et étudions différentes manières d'intégrer les informations fournies par la saillance visuelle au sein du modèle de reconnaissance d'objet *Sac de mots visuels* (*Bag of Visual Words*). D'après nos recherches, il s'agit de la première étude si profondément établie à plusieurs niveaux : i) nous étudions la pondération de caractéristiques par la saillance, ii) nous étendons l'état-de-l'art sur l'échantillonnage non-uniforme des oints caractéristiques d'après les valeurs saillantes, iii) nous introduisons un tout nouveau modèle de codage des caractéristiques basé sur la saillance. Après une discussion sur l'influence de chacun de ces modules et ses paramètres, nous avons montrés comment notre modèle de saillance biologiquement inspiré permet d'étendre les performances du système considéré. Non-seulement obtenons-nous des améliorations notables comparé au système Sac de mots visuels classique mais mais aussi nous fournissons de nouveaux résultats état-de-l'art dans toutes les bases de données égocentriques considérées et ceci pour un coût computationnel très compétitif.

Notre contribution finale a été de démontrer comment la reconnaissance d'activités dans les vidéos égocentriques peut être adressée avec succès en rejoignant deux sources d'information: a) les objets actifs, b) le contexte (basé sur la reconnaissance de lieux). Pour ce résultat, une méthode de reconnaissance d'activités qui modélise celles-ci comme des séquences d'objets actifs et de lieux a été testée sur une base de donnée de vidéos égocentrique non-triviale montrant des scénarios de la vie quotidienne pour plusieurs personnes. Nous avons démontré comment la combinaison d'à la fois objets actifs et contexte fournit des améliorations notables de performance et surpasse les méthodes de l'état-de-l'art utilisant des représentations d'objets actifs + passifs dans les bases de données considérées.

Mots-Clés: Saillance, Reconnaissance d'objets, reconnaissance d'activités, vidéos égocentriques, Sac-de-mots-visuels, Dem@care

Discipline: Informatique

LaBRI (UMR CNRS 5800)
Université Bordeaux 1
351, cours de la Libération
33405 Talence Cedex (FRANCE)

Perceptual Object Of Interest Recognition: Application to the Interpretation of Instrumental Activities Of Daily Living For Dementia Studies

Abstract: The rationale and motivation of this PhD thesis is in the diagnosis, assessment, maintenance and promotion of self-independence of people with dementia in their Instrumental Activities of Daily Living (IADLs). In this context a strong focus is held towards the task of automatically recognizing IADLs. Egocentric video analysis (cameras worn by a person) has recently gained much interest regarding this goal. Indeed recent studies have demonstrated how crucial is the recognition of active objects (manipulated or observed by the person wearing the camera) for the activity recognition task and egocentric videos present the advantage of holding a strong differentiation between active and passive objects (associated to background). One recent approach towards finding active elements in a scene is the incorporation of visual saliency in the object recognition paradigms. Modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered of *interest* or *salient*, which, in egocentric videos, strongly corresponds to the locus of objects of interest. The objective of this thesis is to design an object recognition system that relies on visual saliency-maps to provide more precise object representations, that are robust against background clutter and, therefore, improve the recognition of active object for the IADLs recognition task. This PhD thesis is conducted in the framework of the Dem@care European project.

Regarding the vast field of visual saliency modeling, we investigate and propose a contribution in both *Bottom-up* (gaze driven by stimuli) and *Top-down* (gaze driven by semantics) areas that aim at enhancing the particular task of active object recognition in egocentric video content. Our first contribution on Bottom-up models originates from the fact that observers are attracted by a central stimulus (the center of an image). This biological phenomenon is known as *central bias*. In egocentric videos however this hypothesis does not always hold. We study saliency models with non-central bias geometrical cues. The proposed visual saliency models are trained based on eye fixations of observers and incorporated into spatio-temporal saliency models. When compared to state of the art visual saliency models, the ones we present show promising results as they highlight the necessity of a non-centered geometric saliency cue. For our top-down model contribution we present a probabilistic visual attention model for manipulated object recognition in egocentric video content. Although arms often occlude objects and are usually considered as a burden for many vision systems, they become an asset in our approach, as we extract both global and local features describing their geometric layout and pose, as well as the objects being manipulated. We integrate this information in a probabilistic generative model, provide update equations that automatically compute the model parameters optimizing the likelihood of the data, and design a method to generate maps of visual attention that are later used in an object-recognition framework. This task-driven assessment reveals that the proposed method outperforms the state-of-the-art in object recognition for egocentric video content.

Regarding the challenging task of recognizing active objects we proposed and studied different semantic ways to integrate the information brought by visual saliency modeling into the well-known Bag of Visual Words (BoVW) object recognition paradigm. To the best of our knowledge, this is the first in-depth study about the application of visual saliency to object recognition with BoVW approach at all its stages: i) we study the saliency-based feature weighting, ii) we extend the state-of-the-art on saliency-sensitive non-uniform feature sampling in a new saliency-sensitive variable-resolution feature space, iii) we introduce a completely new saliency-Sensitive Coding of features. After discussing the influence of each module and its parameters, we have shown how our biologically inspired saliency-based model helps to enhance current system performances. It not only achieves notable improvements with respect to the baseline BoVW, but also provides state-of-the-art results in all the considered egocentric datasets at very competitive computational times.

Our final contribution has been to demonstrate how activity recognition in egocentric video can be successfully addressed by the combination of two sources of information: a) active objects either manipulated or observed by the user provide very strong cues about the action, and b) context also contributes with complementary information to the active objects, by identifying the place in which the action is being made. For that end, an activity recognition method that models activities as sequences of active objects and places have been used on a challenging egocentric video dataset showing daily living scenarios for various users. We have demonstrated how the combination of both objects+context provides notable improvements in the performance, and outperforms state-of-the-art methods using active+passive objects representations.

Keywords: Saliency, object recognition, activity recognition, egocentric videos, Bag of Visual Words (BoVW), Dem@care

Discipline: Computer Science

LaBRI (UMR CNRS 5800)
Université Bordeaux 1
351, cours de la Libération
33405 Talence Cedex (FRANCE)

Résumé de la thèse

Cette section présente un résumé substantiel de cette thèse en français. Le contexte et la problématique sont introduits dans cette première partie. La prochaine section présente en détail les objectifs de la thèse. Nous présenterons enfin les différentes contributions de ce travail de thèse.

Context and introduction

Dans le contexte d'une population européenne vieillissante, le développement de nouveaux systèmes et technologies est devenu essentiel afin d'aider les séniors à maintenir leur indépendance et permettre aux médecins de mieux traiter les maladies en lien avec la démence due à l'âge afin de développer de nouveaux traitements thérapeutiques. D'après des études médicales à grande échelle [Peres 08, Helmer 06], les premiers signes de démence peuvent être observés en tant que difficultés fonctionnelles dans les activités de la vie quotidienne jusqu'à 10 ans avant le diagnostique médical défini par les méthodes cognitives de référence [McKhann 84, Hachinski 94]. Les moyens traditionnels d'évaluation de la démence basé sur des questionnaires sont en effet sujets à des biais de différent types tels que la non-acceptation de perte de fonctions cognitives par les patients lors de discussions avec les docteurs ou le visionnage constant et stressant des effets de la démence sur les proches par les aides soignants qui ont tendance à surestimer les difficultés des patients [Helmer 06].

Une méthode d'observation objective des activités de la vie quotidienne (IADLs) directement au domicile des patients pourrait amener des informations complémentaires non biaisées qui pourraient s'avérer très précieuses pour les docteurs en charge du diagnostic afin d'affiner leurs analyses. Le fait de pouvoir observer les patients effectuer leurs activités de la vie de tous les jours dans un environnement familier (leur propre maison par exemple) pourrait permettre en effet d'interpréter correctement les signes précurseurs de démence qui ne seraient pas forcément reportés dans les questionnaires. Afin de permettre au docteurs d'observer et analyser les activités de la vie quotidienne d'un patient, l'utilisation de différents capteurs portés par le patient est une approche originale. Les enregistrements de ces capteurs, si correctement post-traités et présentés, pourraient fournir des informations objectives à propos des potentielles difficultés rencontrées par le patient dans ses activités, pouvant aider le docteur dans son diagnostic afin de permettre de commencer des techniques de rééducation adaptées à un stade précoce.

Ce travail de thèse est mené en partenariat avec le projet européen Dem@care FP7- ICT-2011.5.1 IP conçu spécialement pour ce but spécifique. En effet ce projet a pour vocation de contribuer au diagnostique régulier, l'évaluation, l'entretien et la promotion de l'auto-indépendance des personnes atteintes de démence en approfondissant la compréhension de comment cette maladie affecte leur vie quotidienne. Le projet met en œuvre une boucle fermée multiparamétrique pour les personnes atteintes de démence et leurs aide-soignants afin de surveiller et d'évaluer leur statut cognitif et comportemental en intégrant une multiplicité de capteurs portables et in-situ, afin de permettre une évolution contextuelle de profil des patients pour fournir les soins et informations adaptés.

Parmi ces capteurs portables, nous avons travaillé avec une caméra GoPro monté sur l'épaule. Les caméras portables représentent un moyen abordable et efficace pour enregistrer les activités des patients et offrir un point de vue unique, en particulier sur les objets manipulés, pendant une activité. Ce fait est au cœur de notre recherche. En effet, les objets manipulés (ou avec lesquels il y a interaction) pendant les activités de la vie quotidienne détiennent une quantité incroyable d'informations à leur sujet. Reconnaitre les objets manipulés est donc une première étape logique vers le but qui motive cette thèse qui consiste dans la reconnaissance des activités de la vie quotidienne pour l'évaluation, l'entretien et la promotion de l'auto-indépendance des patients souffrant de la maladie d'Alzheimer et problèmes de démence liés à l'âge.

Objectifs de la thèse

Comme indiqué dans la section précédente ce travail de thèse est motivée par la reconnaissance des activités de la vie quotidienne pour l'évaluation des capacités des patients souffrant de la maladie d'Alzheimer et les problèmes de démence liés à l'âge. La tâche de reconnaître les activités de la vie quotidienne dans les vidéos est devenu un enjeu fondamental entre la communauté de vision par ordinateur [Gaidon 09]. Pour faire face à un champ de vision limité et la difficulté d'accéder à toutes les informations pertinentes à partir de caméras fixes, une alternative a été trouvée dans les vidéos égocentriques, enregistrées par des caméras portées par les sujets. En effet, en plus de traiter avec les inconvénients énumérés précédemment, les caméras portables représentent un moyen abordable et efficace pour enregistrer l'activité des utilisateurs dans des scénarios tels que le nôtre.

Traditionnellement, la détection des activités de la vie quotidienne est adressée par l'analyse des mouvements humain. Plus précisément, diverses approches ont utilisé avec succès le mouvement associé à des points d'intérêt spatio-temporels (STIP) dans la vidéo [Laptev 08, Wong 07]. Toutefois, dans le cas particulier des vidéos égocentriques, nous affirmons que l'action peut être effectivement définie comme une séquence d'objets manipulés ou observés, généralement connu sous le nom d'*objects actifs* ou *objets d'intérêt*. Cette hypothèse est généralement valable pour les vidéos montrant de nombreuses activités domestiques c'est à dire notre cas. Réconfortant cette idée, Pirsiavash et Raman ont récemment démontré dans [Pirsiavash 12] que les performances de reconnaissance sont considérablement améliorées si l'on a connaissance de l'objet manipulé, une idée similaire est explorée dans [Fathi 12].

La reconnaissance d'objet dans la vidéo est un problème ouvert, indépendamment de la nature du contenu. Contrairement aux approches par fenêtre coulissante bien connu détection et la reconnaissance d'objets [Felzenszwalb 10, Lampert 08], et en raison de la nature spécifique de la vue à la première personne que nous considérons, nous visons à guider le processus de reconnaissance d'objets vers des zones d'intérêt à l'aide saillance visuelle.

La saillance visuelle dans une image est définie par les régions ou les détails qui attirent le regard. La modélisation du processus sélectif du système visuel humain représente un moyen efficace pour guider l'analyse de la scène vers les endroits *d'intérêt* ou *saillant*. C'est une des raisons pour lesquelles la saillance visuelle est devenue une tendance très active dans le domaine de la vision par ordinateur [Borji 12a]. Grâce à l'utilisation de cartes de

saillance, la recherche d'objets dans des images est plus ciblée, améliorant ainsi la performance de reconnaissance tout en réduisant en la charge de calcul [Sharma 12, San Biagio 14]. L'incorporation de la saillance visuelle dans l'étude de contenu vidéo est une tendance récente. De même que pour les images fixes, l'application de la modélisation de saillance pour la reconnaissance d'objet dans la vidéo permet d'identifier les zones où se trouvent les objets d'intérêt. Pour le scénario particulier des vidéos égocentriques, il existe généralement une forte différenciation entre les objets actifs (manipulés ou observés par l'utilisateur portant la caméra) et passifs (associés à l'arrière plan), par conséquent des indices sémantiques spatiaux, temporels, ou géométriques peuvent être extraits du contenu vidéo afin de modéliser la saillance visuelle pour identifier les éléments actifs de la scène. Plusieurs œuvres de la littérature ont montré l'utilité de suivi du regard humain dans l'analyse de contenu vidéo égocentrique et, en particulier, dans la tâche de reconnaissance d'activité [Fathi 12, Ogaki 12, Vig 12].

C'est notre conviction que la reconnaissance d'objets *actifs* est une première étape essentielle pour la reconnaissance d'activités de la vie quotidienne, c'est pourquoi dans cette thèse nous étudions la combinaison de deux domaines essentiels dans la vision par ordinateur: la reconnaissance d'objets et la saillance visuelle. Plus précisément, nous présentons et étudions des modèles de reconnaissance d'objets bénéficiant de la puissance discriminative des cartes de saillance afin de reconnaître seulement et avec plus de précision les objets actifs dans les vidéos égocentriques.

Contributions de la thèse

Nos premières contributions concernent le domaine de la modélisation de saillance visuelle. En effet, nous proposons et étudions une contribution à la fois dans le domaine bottom-up (regard entraîné par un stimulus) et top-down (regard entraîné par la sémantique) visant à renforcer la tâche particulière de reconnaissance d'objet actif dans le contenu vidéo égocentrique. Concernant les modèles bottom-up, notre première contribution provient du fait que les observateurs sont généralement attirés par un stimulus central (le centre de l'image). Ce phénomène biologique est connu sous le nom de *biais central*. Dans les vidéos égocentriques cette hypothèse n'est cependant pas toujours vérifiée. Pour cette raison, nous proposons d'étudier des modèles de saillance avec des biais non-centraux. Les modèles de saillance visuelle proposés sont basés sur les fixations oculaires d'observateurs et incorporés dans les modèles de saillance spatio-temporelles. Par rapport aux modèles de saillance visuelle de l'état-de-l'art, ceux que nous présentons montrent des résultats prometteurs car ils mettent en évidence la nécessité d'une saillance géométrique non-centrée. Dans notre deuxième contribution, nous présentons un modèle probabiliste de l'attention visuelle top-down pour la reconnaissance d'objet manipulé dans le contenu vidéo égocentrique. Contrairement à la plupart des modèles top-down existants [Ma 05, Cerf 07, Pinto 13], nous n'avons pas besoin d'entrainer des détecteurs d'objets préalables dans ce modèle, nous utilisons seulement les informations constamment fournies par les bras et les mains. En effet, bien que les bras sont sources d'occlusion des objets et généralement considérés comme un fardeau pour de nombreux systèmes de vision, ils deviennent un atout dans notre approche puisque nous ex-

trayons des caractéristiques décrivant leur disposition géométrique et intégrons cette information dans un modèle génératif probabiliste optimisé pour générer des cartes d'attention visuelle qui sont ensuite utilisées dans un modèle de reconnaissance d'objet. Des évaluations rigoureuses montrent que cette méthode surpasse l'état-de-l'art en reconnaissance d'objets manipulés dans les vidéos égocentriques.

En ce qui concerne la tâche de reconnaître les objets actifs, nous avons proposé et étudié différentes façons sémantiques pour intégrer l'information apportée par saillance visuelle dans le modèle de reconnaissance d'objet Sac-de-mots-visuels (BoVW). D'après nos recherches, ceci est la première étude approfondie sur l'application de la saillance visuelle pour la reconnaissance d'objets avec l'approche BoVW à tous ses stades:

1. nous étudions la pondération de la signature basée sur la saillance,
2. nous étendons l'état-of-the-art sur l'échantillonnage non-uniforme des caractéristiques basé sur la saillance grâce à une nouvelle fonctionnalité d'extraction de caractéristiques à résolution variable basé sur la saillance.
3. nous introduisons un nouveau codage des caractéristiques basé sur la saillance.

Après avoir discuté de l'influence de chaque méthode et ses paramètres, nous avons montré comment notre modèle biologiquement inspiré contribue à améliorer les performances des systèmes actuels. Il apporte non seulement des améliorations notables par rapport au modèle Sac-de-mots-visuels, mais fournit également les résultats de l'état-de-l'art dans toutes les bases de données égocentriques considérés pour des temps de calcul très compétitifs.

Puisque nous cherchons à reconnaître des objets actifs pour la reconnaissance des activités de la vie quotidienne, notre contribution finale a été d'étudier comment la reconnaissance d'activités dans les vidéos égocentriques peut être traitée avec succès par la combinaison de deux sources d'information:

- les objets actifs
- l'emplacement dans lequel l'activité se passe.

Dans cette optique, un modèle de reconnaissance d'activités qui considère les activités comme des séquences d'objets actifs et les lieux a été utilisé sur un ensemble de données vidéos égocentriques complexes montrant des scénarios de la vie quotidienne pour différents utilisateurs. Nous avons démontré comment la combinaison d'objets actifs + contexte offre des améliorations notables dans les performances de reconnaissance, et surpasse les méthodes de l'état-de-l'art.

Contents

List of Figures

List of Tables

Acronyms

AP Average Precision. xv, xvii, 59, 61, 63, 80–82, 112

AUC Area Under Curve. 48, 49

BB Bounding Box. xvii, 56, 59, 60, 63, 79

BoVW Bag of Visual Words. xiv, xv, 5, 23, 25, 27, 28, 31, 33, 34, 59, 63, 64, 76, 79, 82, 85–87, 92, 97, 99–101, 105, 109, 119–122

BoW Bag of Words. xiii, 24, 25, 85, 99–105, 110

CNN Convolutional Neural Networks. 30, 33

Dem@care Demantia Ambient Care. iii, 1, 119

DNN Deep Learning Networks. 45

DoG Difference-of-Gaussian. 13, 15

DPM Deformable Part-Based Model. xiii, 28–30, 79, 101, 102, 104, 116

DPMs Deformable Part-Based Models. 28, 30, 31

FIT Feature Integration Theory. 35, 37, 40, 41, 44

FoA Focus of Attention. 41, 44

GBVS Graph-Based Visual Saliency. xvii, 44, 57–63, 72, 75, 76, 80, 81

GME Global Motion Estimation. 41, 53

GT Ground Truth. 99, 100, 102–104

HCI Human Computer Interactions. 7

HOF Histogram of Optical Flow. 17

HOG Histogram of Oriented Gradients. 17

HVS Human Visual System. xiii, 35, 36, 41, 85, 122

IADL Instrumental Activity of Daily Living. xi, 4, 5, 119, 120, 122

IADLs Instrumental Activities of Daily Living. xi, 1–3, 37, 107, 109, 110, 113, 119

IMS Intégration du Matériau au Système. iii, 110

KLB Kullback-Leiber divergence. 48

LaBRI Laboratoire Bordelais de Recherche en Informatique. iii

LBP Local binary patterns. 16

LLC Locality-constrained Linear Coding. 93, 98

LoG Laplacian-of-Gaussian. xiii, 13–15

mAP Mean Average Precision. xvi, xvii, 32, 59, 61–64, 76, 77, 79–81, 97, 100, 102, 112, 117

NSS Normalized Scanpath Saliency. xvii, 48–50, 74, 76

NUS Non Uniform Sampling. xv, xvii, 90, 91, 97, 98, 100, 101

PCC Pearson Correlation Coefficient. xvii, 48, 59–61

PDF probability density function. 27

RANSAC Random Sample Consensus. 20, 21, 53

SC Saliency-Sensitive Coding. x, xv, xvii, 86, 92, 93, 98–101, 104

SIFT Scale-Invariant Feature Transform. xiii, 13, 16–18, 20, 22, 25, 87

SP Saliency-based Pooling. x, 87, 100, 101

SS Segmentation as Selective Search. 32, 101, 102

STC Spatio-Temporal geometric Centered. xvii, 54, 57–63, 74–76, 80, 81

STNC Spatio-Temporal geometric Non-Centered. xiv, 54, 56–64

STNCF Spatio-Temporal geometric Non-Centered Flared. xiv, 54, 56–64

SURF Speeded Up Robust Features. xiii, 17, 19, 25, 59, 87, 102

SVM Support Vector Machine. 30, 32, 34, 44, 59, 76, 87, 110–112

tf-idf Term Frequency-Inverse Document Frequency. 24

VSR Variable Resolution Sampling. xv, xvii, 86, 90, 97, 98, 100, 101

WTA Winner Take All. 41, 44

Chapter 1

Main introduction

In the context of an aging European population, developing new home care technologies and systems has become crucial for helping seniors maintaining their independence and allowing medical practitioners to better study age-related dementia in order to come up with adapted therapeutic treatments. Large scale medical studies [Peres 08, Helmer 06] have shown that earliest signs of dementia can be observed as functional difficulties in daily living activities up to 10 years before the clinical diagnostic defined by the cognitive methods of reference [McKhann 84, Hachinski 94]. Indeed, the traditional ways of assessment of dementia based on questionnaires are prone to different types of bias such are the non-acceptance of loss of cognitive functions by the patients when reporting to doctors or the constant stressful witnessing of the effects of dementia on relatives by the caregivers which tend to over estimate the difficulties of patients with dementia[Helmer 06].

An objective observation method at the patient's home of its Instrumental Activities of Daily Living (IADLs) could potentially bring additional and unbiased information that would be valuable for the doctors in charge of the diagnosis to further refine their analysis. The observation of the patients performing everyday's activities in their ecological and familiar environments at home would indeed allow a correct interpretation of the early signs of dementia that might not be reported in the questionnaires. In order to enable the doctors to observe and analyze one patient's daily living, the use of several sensors worn by the patient is an original approach. The sensors recordings of daily living activities, if post-processed and presented accordingly, would give an objective input about potential difficulties in the activities, helping the doctor in his diagnosis, enabling the setup of adapted reeducation techniques and the evaluation of therapeutic efficiency.

This PhD work is conducted in the framework the Dem@care FP7-ICT-2011.5.1 IP European project designed for this specific goal. This project aspires to contribute to the timely diagnosis, assessment, maintenance and promotion of self-independence of people with dementia, by deepening the understanding of how the disease affects their everyday life and

Figure 1.1 – The recording device (GoPro camera) fixed on the vest offering an egocentric point of view

behavior. It implements a multi-parametric closed-loop for people with dementia and their informal caregivers to monitor and assess their cognitive and behavioral status by integrating a multiplicity of wearable and in-situ sensors, enable time evolving context-sensitive profiling to support reactive and proactive care, and afford personalized and adaptive feedback.

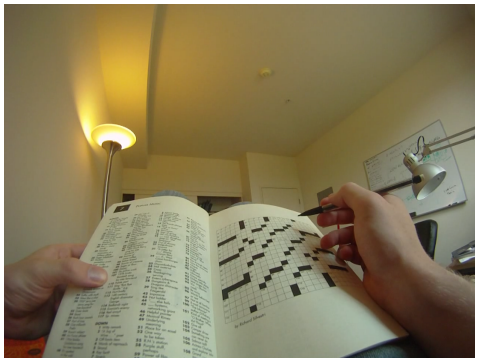
Among these wearable sensors, we specifically came to work with a shoulder-mounted GoPro camera. Wearable cameras represent a cheap and effective way to record patients' activities, and offer a unique point of view, especially on the objects being manipulated during an activity (see Figures 1.1 and 1.2). This fact is at the heart of our research. Indeed, manipulated objects (or interacted with) during IADLs hold an incredible amount of information about them. Finding manipulated objects is therefore a first and logical step towards the higher goal motivating this whole thesis which consist in the recognition of IADLs for the assessment , maintenance and promotion of self-independence of patients suffering from Alzheimer disease and age-related dementia issues.

In this chapter we will first detail the thesis objectives and especially how the recognition of instrumental activities can be addressed with egocentric video content for healthcare and specifically for dementia studies. We will then present and describe the various contributions of this PhD work and will finish by providing the outline of the manuscript.

1.1 Thesis objectives

As stated in the previous section this thesis work is motivated by the recognition of IADLs for the assessment of the ability of patients suffering from Alzheimer's disease and age-related dementia issues. The task of recognizing human activities in videos has become a fundamental challenge among the computer vision community [Gaidon 09]. In order to face the limited field of view and the difficulty of accessing all relevant information from fixed cameras, an alternative has been found in egocentric videos, recorded by cameras worn

(a) GTEA Dataset [Fathi 11b] (b) EDSHK dataset [Li 13b]



(c) ADL dataset [Pirsiavash 12]

Figure 1.2 – Examples of egocentric datasets illustrating the unique point of view on manipulated objects.

by subjects. Indeed, in addition to dealing with the previously listed drawbacks, wearable cameras represent a cheap and effective way to record users activity for scenarios such as ours.

One of the first project involving the use of wearable imaging device is the SENSECAM project [Hodges 06], which aims to provide wearable image lifelog as a memory aid. The camera is worn around the neck and captures pictures at several seconds of interval during the day. Finally, the events of the day are summarized as automatic life-logs [Berry 07]. A wearable camera has also been used in the WearCam project [Piccardi 07], where the camera is mounted on the head of children to help early diagnosis of autism. The automatic analysis of the child gaze during the execution of specific movements is interpreted for the diagnosis. More recently, egocentric videos have been used for the recording of IADLs on patients with dementia in the ANR IMMED project [Mégret 10]. First promising results of recognition of IADLs in such recorded video were reported in [Karaman 14]. Nevertheless, the recognition problem being very complex, efficient ways of solving it still remain an open research issue.

Traditionally, the detection of human activities has been addressed by analyzing human motion patterns. More precisely, various approaches have successfully made use of the motion patterns associated to spatio-temporal interest points (STIP) in the video [Laptev 08, Wong 07]. In addition, the study of ego-motion has also resulted in successful approaches for first-person camera videos analysis [Kitani 11]. However, in the particular case of egocentric view, we claim that an action can be effectively defined as a sequence of

manipulated or observed objects, usually known as *active* objects or *objects-of-interest*. This assumption generally holds for video showing many household activities and, in particular, for the intended IADL scenario. Comforting this idea, Pirsivash and Ramanan have recently demonstrated in [Pirsiavash 12] that performances are dramatically increased if one has knowledge of the object being interacted with, a similar idea is explored in [Fathi 12].

Visual object recognition in video is an open problem regardless from the nature of the content. In contrast to the well-known sliding window approaches for object detection and recognition [Felzenszwalb 10, Lampert 08], and due to the specific nature of the first-person view content, we aim to drive the object recognition process towards relevant zones using visual saliency.

Visual saliency in an image is defined by the regions or details that attract human gaze. Modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered *of interest* or *salient*. This is one of the reasons why it has become a very active trend in computer vision [Borji 12a]. Thanks to the use of saliency maps, the search for objects in images is more focused, thus improving the recognition performance and additionally reducing the computational burden [Sharma 12, San Biagio 14]. The incorporation of visual saliency in video content understanding is a recent trend. Similarly to still images, the application of saliency modeling for object recognition in video helps identifying areas where objects of interest are located. Under the particular scenario of egocentric videos, there is usually a strong differentiation between active (manipulated or observed by the user wearing the camera) and passive objects (associated to background) and, therefore, spatial, temporal, geometric or even semantic cues can be found in the video content that may help modeling visual saliency to identify the active elements in the scene. Several works in the literature have shown the utility of human gaze tracking in the analysis of egocentric video content and, in particular, in the activity recognition task [Fathi 12, Ogaki 12, Vig 12].

It is our belief that *active object recognition* is a first and essential step for the IADL recognition task, therefore in this thesis we wish to study the combination of two important fields in computer vision: object recognition and visual saliency. More specifically we present and study models of objects recognition benefiting from the discriminative power of saliency maps in order to recognize solemnly and more accurately active objects in egocentric videos.

1.2 Thesis contributions

Our first contributions concern the field of visual saliency modeling. Indeed we investigate and propose a contribution in both Bottom-up (gaze driven by stimulus) and Top-down (gaze driven by semantics) areas aiming at enhancing the particular task of active object recognition in egocentric video content. Concerning the Bottom-up models, our first contribution stems from the fact that observers are usually attracted by a central stimulus (the center of an image). This biological phenomenon is known as *central bias*. In egocentric videos however this hypothesis does not always hold. We study saliency models with non-central bias geometrical cues. The proposed visual saliency models are trained based on

eye fixations of observers and incorporated into spatio-temporal saliency models. When compared to state-of-the-art visual saliency models, the ones we present show promising results as they highlight the necessity of a non-centered geometric saliency cue. In our second contribution, we present a probabilistic Top-down visual attention model for manipulated object recognition in egocentric video content. Unlike most existing Top-down models [Ma 05, Cerf 07, Pinto 13], we do not need to train prior object detectors of any kind in this model but use the information constantly provided by arms and hands. Indeed, although arms often occlude objects and are usually seen as a burden for many vision systems, they become an asset in our approach, as we extract features describing their geometric layout and pose and integrate this information in an optimized probabilistic generative model to generate maps of visual attention that are later used in an object-recognition framework. This task-driven assessment reveals that the proposed method outperforms the state of the art for the manipulated object recognition in egocentric videos.

Regarding the challenging task of recognizing active objects we proposed and studied different semantic ways to integrate the information brought by visual saliency modeling into the well-known BoVW object recognition paradigm. To the best of our knowledge, this is the first in-depth study about the application of visual saliency to object recognition with a BoVW approach at all its stages: i) we study the Saliency-based feature weighting, ii) we extend the state-of-the-art on Saliency-sensitive non-uniform feature sampling in a new Saliency-sensitive variable-resolution feature space, iii) we introduce a completely new Saliency-Sensitive Coding of features. After discussing the influence of each method and its parameters, we have shown how our biologically inspired saliency-based model helps to enhance current system performances. It not only achieves notable improvements with respect to the baseline BoVW, but also provides state-of-the-art results in all the considered egocentric datasets at very competitive computational times.

Since we aim at recognizing active objects for the IADL recognition task, our final contribution has been to study how activity recognition in egocentric video can be successfully addressed by the combination of two sources of information: a) active objects and b) location in which the activity is happening. For that end, an activity recognition framework that models activities as sequences of active objects and places have been used on a challenging egocentric video dataset showing daily living scenarios for various users. We have demonstrated how the combination of both objects+context provides notable improvements in the performance, and outperforms state-of-the-art methods using active+passive objects representations.

1.3 Thesis outline

Here is presented the organization of the manuscript.

First of all, this work tackles two major domains of computer vision: how to detect objects and how to model visual attention. Both fields of research, on their owns, are non-trivial and have been at the origin of countless new ideas and breakthroughs. It is therefore necessary for this manuscript to start by presenting both fields in separate state-of-the-art chapters in

order to provide the necessary insights before elaborating our contributions.

Hence the next chapter 2 will be dedicated to the object recognition problem, where the main concepts will be presented: how to extract and match descriptors, how to use them to build insightful and discriminative representations of pictures and how to classify these representations.

We will then review the state-of-the-art in visual saliency modeling in chapter 3. In particular we will start by discussing the characteristics and definition, present the features for automatic saliency modeling, describe the techniques to build visual attention maps from human eye fixations and finally present the metrics to evaluate saliency maps depending on the task they are built for.

Chapter 4 brings our first contributions regarding the modeling of visual saliency. We present both a new bottom-up and top-down model and evaluate their performances for the particular task of manipulated object detection.

In chapter 5 we provide an in-depth exploration and new ways to combine saliency maps with object recognition systems with BoVW approach at all its stages, showing how much biologically inspired saliency-based model allows to enhance current system performances.

We present our final contribution in chapter 6 where we demonstrate how activity recognition in egocentric video can be successfully addressed by the combination of two sources of information: manipulated objects and spatial information (for the context).

Finally, chapter 7 gives the general conclusions and describes the ongoing work as well as future perspectives.

Chapter 2

State of the art in Object Recognition

The advantages brought by accurate object recognition systems can be beneficial to many sectors of research or industry such as Human Computer Interactions (HCI), surveillance, biometrics, robotics, health care, ...

In the commonly addressed *object recognition problem*, the word *object* has a large definition (object, activity, gesture, ...), that is why this field of research is also usually called pattern recognition. Before getting any further, it is important to stress out the difference between the so-called "object categories" and "object instances". Instances refer to one object in particular, with no room for variations, while an object category refers to the kind of object and has room for a lot of variability (colors, shape, ..., see Figure 2.1).

Generally adults are able to recognize more than 10 000 categories of objects [Biederman 95], and an even much higher quantity of instances. Moreover the recognition process is fast, almost effortless, and robust to variations of viewpoints, luminosity and occlusions. Given only very few images of an object, a person can learn a new object category very quickly and it requires only one picture to learn a new instance.

Many neuroscientific studies have tried to understand the functioning of our brain when it comes to visual recognition. The authors of [Schwartz 77] have discovered that our brain contains several layers of neurons dedicated to different low-level processing of the information coming from the eye. Those layers contain different neuron types called C1, V1, C2 and V2 which are known to apply some simple fixed pre-processing, such as extracting local edges or gradient orientations and aggregating those information.

Interestingly enough, object recognition systems roughly follow the same dataflow (see Figure 2.2). Firstly low-level image features such as edges, corners, textures,... are extracted from the images. The second step consists in learning amendable models of objects from these extracted features, allowing to effectively take a decision about the presence of a given model object in a test image. Finally, a more complex decision process, previously trained to

(a) 4 representations of the same instance: a painting of the Joconde

(b) 4 objects belonging to the same category: painting

Figure 2.1 – Illustration of the difference between object instances and categories

distinguish between the model objects, is run in a last step.

In the rest of this chapter, we begin by presenting most of the popular existing feature detectors and descriptors. In a second part this chapter will introduce matching techniques and their applications for instance recognition problems. Then, we present ways to aggregate these low-level information in feature-spaces more fitted for class-object detection. Finally we give an overview of existing recognition methods employed in class object recognition and introduce how visual saliency (presented in the next chapter) has been combined with class-object recognition paradigms. For the sake of clarity and space and due to the extensive amount of methods for visual object recognition, some aspects might be less developed than others in this state-of-the-art. The reader is referred to the very thorough following works for complimentary details: [Grauman 11, Andreopoulos 13]

2.1 First step: extracting low-level features

As mentioned previously, the usual first step in object recognition is to extract low-level features from the images, as an intermediary step before more complex processing. Features are simply values computed from pixels according to specific operations. Their aim is to carry the most possible information about objects while being as discriminating and invariant as possible.

We can distinguish between two families of features. *Global features*, as their name suggests, originates from the whole image, for example the distribution of colors in an image. The other type are *local features* which, oppositely, are only computed on a limited area of the image.

Throughout this work, we used only local features as they are more fitted for the object

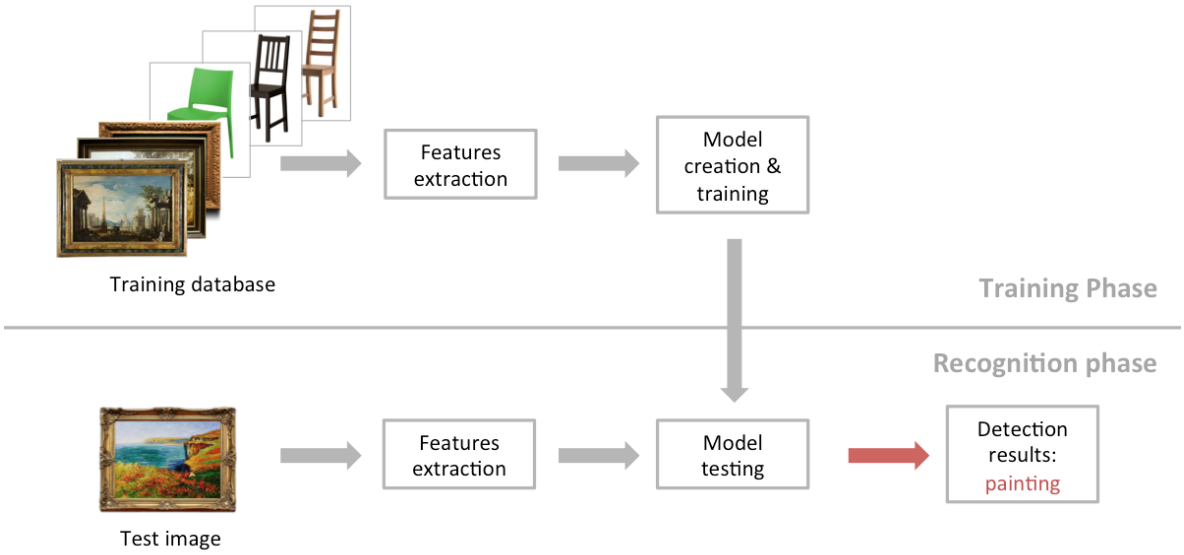


Figure 2.2 – Illustration of the usual dataflow for object recognition systems

category recognition task. Indeed they are designed to be computed on regions corresponding to the objects themselves and therefore are invariant to translations. They also have many other properties that will be presented later. Global features on the other hand are more inclined to tasks such as scene classification (whether a picture is taken in the sea or in a forest) since they consider the image as a whole.

In this section we will start by talking about global features, then we will introduce the local ones and describe more deeply certain well-known methods.

2.1.1 Global features

Global image features rely on the image on its whole to create a signature. Among the first and most famous global descriptors are the ones using the color information present in images such as color histograms and color moments.

Color histograms represent the distribution of the colors present in the image (see Figure 2.3). Each bin is associated to the frequency of a color value within the image. Histograms present the advantage to be invariant under geometrical transformations and some tolerance for partial occlusions. They were first introduced for the task of instance recognition by [Swain 91]. More recent works have introduced color moments as another way of representing the color distribution of an image [Jing 02, Long 03]. The first order moment compute the average color value of the image, the second order moment gives information about the variance of the color values and the third order moment computes the asymmetry degree of the color distribution, called skewness. Other color descriptors that can be mentioned are the Dominant Color Descriptor (DCD) introduced in the MPEG-7 standard or the Color

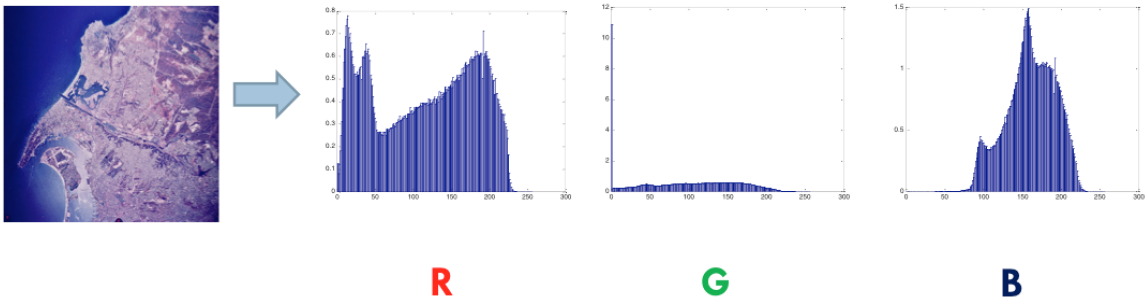


Figure 2.3 – Illustration of color histograms along the Red, Green and Blue channels

Layout Descriptor (CLD).

Among the first statistical models proposed using global features is the Eigenfaces approach [Sirovich 87, Turk 91]. This approach was the first one used with success in for the task of face recognition. It is based on the decomposition of the image in a vector-space derived from the eigenvectors of the covariance matrix built from the concatenated training images. Later, Belhumeur and Kriegman [Belhumeur 96] proposed to optimize the class separability by working in a subspace obtained by Fisher's Linear Discriminant Analysis. The resulting Fisherfaces approach achieves better discrimination capabilities and can be applied to construct subspaces optimized for specific discrimination tasks (such as distinguishing people wearing glasses from people wearing no glasses).

Global methods are robust to modifications of lighting and contrast as long as the set of reference images holds a high enough number of data. However, since the signature is built from all pixels, global features perform poorly if the background is changed, the objects are deformable or if there are occlusions. Another main limitation of these approaches is their requirement that the objects must be closely cropped and aligned.

2.1.2 Local features

Local features address the principal issues of global ones. Images are considered as a collection of local regions instead of one whole entity. The regions generally consist of small round or rectangular patches of an image around a center point. The advantages of working with such small patches are that the features overcome most difficulties encountered during the processing of entire images since they can be standardized. Indeed local regions generally are located such that their scale and orientation can be brought back to standard values. Moreover changes of illumination are way less probable and less important on such a small regions, allowing contrast and luminosity to be normalized. Another advantage is that these regions, when located inside objects, are independent to background changes. Finally the occlusions are less of a problem since there are less chances for partial occlusions: usually local features are either not or totally occluded but they appearances do not change because of occlusions. Because of all these advantages and their computational speed (for most), local features are more fitted for the object recognition task and have been employed more frequently than global ones by the computer vision community.

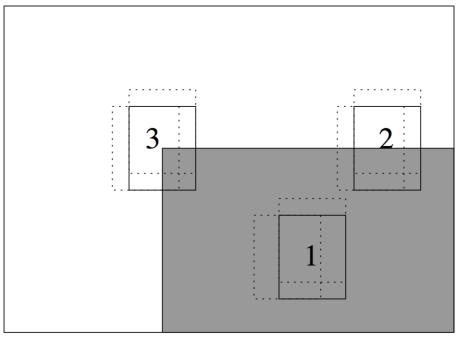


Figure 2.4 – Different main situations in the Moravec corner detection scheme

The determination of the location of these local regions gives two types of local features: the dense and sparse ones. In the case of sparse features, a preliminary step is necessary to compute where the patches locations are to be extracted (usually selected in a way that is invariant to common transformations such as rotation, translation,...). This is called the *keypoint detection* phase. For dense features however, there are no such constraints and the keypoints can be extracted anywhere on the image (usually on a *dense grid*, which corresponds to a squared spatial sampling). This first part of building local features concerns mainly their location, the second part is common to both dense and sparse features: it is the *descriptors construction*.

In the next section (2.1.2.1) we explain how keypoints are detected for building sparse features. Since the construction of descriptors is independent from seeking their location, the descriptors construction methods for dense or sparse local features are presented in section 2.1.2.2. This section is of course a summary of the basic ideas behind keypoints detection and description, the reader can be referred to the work of [Tuytelaars 08] for more details and comparisons.

2.1.2.1 Keypoints localization

The goal of this first step towards building sparse local features is to find locations that can effectively be found again by the same procedure under different viewpoints (translation or rotation), illumination conditions or noise. These locations are called keypoints. Below we present some well-known procedures to localize such keypoints.

First works Most researchers agree that the work on keypoints localization was initiated by Moravec [Moravec 80] and his corner detection algorithm. A corner is defined as a point with low self-similarity. For a given pixel location (u, v), the algorithm considers a patch of adjacent pixels and compares it to other nearby overlapping patches (see fig 2.4). The corner detection function $E(x, y)$ is computed by taking the sum of squared differences (SSD)

between the two patches.

$$E(x,y) = \sum_u \sum_v w(u,v)(I(x+u, y+u) - I(u,v))^2 \tag{2.1}$$

where $w(u,v)$ is the neighbor patch considered (value of 1 inside the patch and 0 outside), $I(u,v)$ is the intensity at pixel (u,v). Figure 2.4 illustrates the three possible detection cases.

1. case 1: if the pixel is a region with uniform intensity values, E will be have low values for all x and y values.
2. case 2: if the pixel is on an edge, the value of E will be small along the edge and high in the direction perpendicular to it.
3. case 3: if the pixel is located on a corner then none of the nearby patches will look similar and the corner detection function will have a high value for any values of x and y.

Consequently the aim of the Moravec corner detector is to search local maximas of the minimum values of E (above a certain threshold). One of the main drawbacks with this operator is that it is not performing uniformly in all orientations (non isotropic). Indeed the neighbor patches can only have discrete values (step of 45˚), hence if an edge is present but not horizontal, vertical or diagonal, the detector will detect a corner since the values of E will be high for all values of x and y.

This detector was improved by Harris and Stephens to better discriminate corners from edges [Harris 88]. The basic idea is based on the first order derivatives of the image.

If we approximate the term $I(x+u, y+u)$ in equation 2.1 by its first order Taylor expansion we obtain:

$$I(x+u, y+u) \approx I(u,v) + I_x(u,v)x + I_y(u,v)y \tag{2.2}$$

Considering this approximation, now equation 2.1 becomes:

$$E(x,y) \approx \sum_u \sum_v w(u,v)(I_x(u,v)x + I_y(u,v)y)^2 \tag{2.3}$$

This can be rewritten in a matrix form:

$$E(x,y) \approx \begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.4}$$

with:

$$A = \sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_y I_x \rangle & \langle I_y^2 \rangle \end{bmatrix} \tag{2.5}$$

Where the brackets mean the weighted summation over all u, v by $w(u,v)$. In this expression the A matrix is a Harris matrix. If the window $w(u,v)$ is circularly weighted (such as a Gaussian), then the response is isotropic.

As for the Moravec detector, a corner can be detected for high values of $E(x, y)$. Harris and Stephens observe that this property can be expressed by observing the magnitudes of eigenvalues λ_1 and λ_2 of the matrix A:

- if both λ_1 and λ_2 are very low (close to 0) then there is nothing to detect at pixel (x, y)
- if λ_1 is low and λ_2 is high (or the opposite), then an edge is found.
- if both λ_1 and λ_2 have high values then a corner is found

Since computing eigenvalues is computationally expensive, Harris an Stephens suggest to compute the following value instead:

$$M = \lambda_1\lambda_1 - \kappa(\lambda_1 + \lambda_2)^2 = det(A) - \kappa.trace^2(A) \tag{2.6}$$

where κ is a parameter to control the sensitivity of detection of interesting point. If the value of M is higher than a threshold t, then an interest point is detected, and therefore the eigenvalues do not need to be computed but simply the trace and determinant of A.

Harris's detector was first employed to image matching problems. The work in [Schmid 97] extend Harris' detector to the general recognition problem for image retrieval. It had been widely used in the community since then due to the fact it can produce repeatable corners.

Scale invariance Harris' detector has proved to be very robust to illumination changes, noise and translations/rotations [Schmid 00] however keypoints are not detected anymore if the scale of the object changes too much. In order to achieve scale invariance, the basic idea would be to compare each probable keypoint location to neighborhoods resized at different scales. Obviously this idea is not feasible because too computationally expensive. Instead it is common measure to start by building up a scale-space [Witkin 83], that is to say a collection of the same image convoluted with a Gaussian Kernel $G(x, y, \sigma)$ at different scales σ.

This idea was extended by Lindeberg [Lindeberg 98] who proposed to build a scale space as a collection of image responses produced by the application of LoG at different normalized scales. Due to the specific form of the LoG (see bottom left of figure 2.5), this proposed detector was more fitted to detect blob-like features. With this detector a keypoint location can be defined as the blob center of a *scale-space extrema*, i.e. the points that are simultaneously local maxima with respect to both space and scale, making it possible to detect both the location+scale of the keypoints.

One of the contribution of David Lowe in his well know SIFT descriptor [Lowe 04] was to show that the Laplacian-of-Gaussian could be approximated by the Difference-of-Gaussian (DoG) $D(\mathbf{x}, \sigma)$. This is way more efficient than computing the derivatives in the case of the LoG since DoG consists instead in a simple difference between two adjacent Gaussian scale spaces separated by a factor k.

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \star I(x, y) \tag{2.7}$$

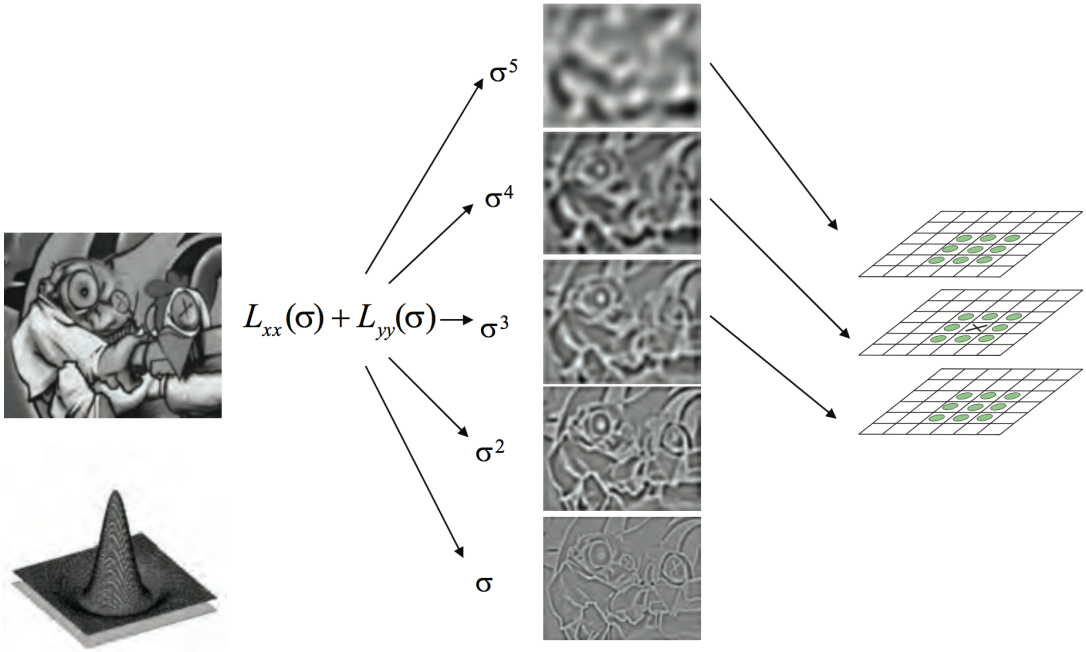


Figure 2.5 – The LoG detector searches for 3D scale space extrema of the LoG function, source: [Tuytelaars 08]

Figure 2.6 – The Laplacian-of-Gaussian can be approximated as a difference of two Gaussian smoothed images, source: [Tuytelaars 08]

Lowe creates scale octaves as a collection of images having the same size and convoluted with different values $k\sigma$ such that $k = 2^{(\frac{1}{K})}$, $\sigma_n = k^n\sigma_0$, and K is a defined number of intervals.

DoG extrema are found by comparing the $D(x, y, \sigma)$ value of each point with its 8-neighbors at the same scale level, and with the 9 closest neighbors on the lower and higher scales. Generally this extremum detection phase produces a lot of keypoints candidates, moreover their location is inaccurate, especially for those points extracted from high octaves (low resolution). That is why post-processing treatments are applied, first to re-estimate correctly the locations, but also to eliminate points with low contrasts or located on edges. In practice the obtained regions are very similar to those extracted with the LoG detector, the DoG detector is therefore often the preferred choice, since it can be computed far more efficiently.

Among other scale invariant detectors worth presenting is the Harris-Laplacian detector [Mikolajczyk 04] which was designed and has proved to be highly discriminative. It combines the LoG scale-space presented in [Lindeberg 98] with the corner detection abilities of the Harris detector and return points which are detected by both methods. This high discriminative power can be a drawback however since for many object recognition applications, the lower number of keypoints might be a disadvantage (especially against occlusions).

Orientation invariance Keypoints should be extracted at the same location despite the possible rotation of an object. One way to achieve orientation invariance is to consider the dominant orientation in a region around the keypoint and rotate it back by an oppositely equal angle. In [Lowe 04], Lowe proposes to consider a grid of pixels around the keypoint which size depends on the corresponding scale σ. Then an histogram of gradients is built with 36 bins representing a full 360°spectrum. When the value of a pixel's corresponding gradient is assigned to a bin, it is weighted by a Gaussian kernel centered on the keypoint location with a scale of 1.5σ so that the center pixels have more impact. The highest bin is selected, interpolated with its two adjacent bins and defined as the main orientation. If the second highest bin has a value higher than 80% the first one, Lowe suggests to add a new

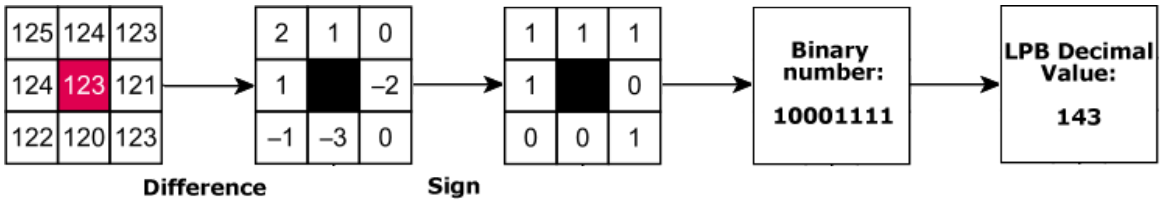


Figure 2.7 – Illustration of the computation of local binary patterns

keypoint at the same location bearing this new orientation since selecting only one could lead to missed matches. It was proven in [Lowe 04] to significantly improve the performances at the cost of only around 15% more keypoints.

2.1.2.2 Descriptors

Once a set of interest points has been extracted from an image, the local neighborhood around each point is encoded in a descriptor designed to be at the same time small in dimensions and highly discriminative. We will present here some of the popular local descriptors.

Local binary patterns (LBP) LBP have originally been introduced by Ojala et al. [Ojala 96] for texture classification, but has been extended to various applications such as face recognition [Ahonen 06], facial expressions recognition [Zhao 07] or human detection [Mu 08]. Their robustness to illumination changes and fast computation time made them highly successful.

The general principle is to compare the value of a pixel with its neighbors. It gives an information about regular patterns in an image, that is to say the texture. More specifically, N pixels are selected on circle around the keypoint (corresponding to the neighborhood with radius R). Each of these pixels gives an output of 1 if higher or equal than the keypoint value, and 0 otherwise. A binary number is formed as the series of these outputs.

$$LBP_{R,N} = \sum_{i=0}^{N-1} s(n_i - n_c)2^i, s(x) = \begin{cases} 0, & x \geq 0 \\ 1, & \text{otherwise} \end{cases} \tag{2.8}$$

where n_c is the gray-scale value at the keypoint location and n_i is the gray-scale value for N pixels equally sampled on the circle with radius R. An illustration of the general principle of LBP is given in Figure 2.7.

The Scale-Invariant Feature Transform (SIFT) descriptor It is in 1999 that David Lowe presents a new method for the extraction, description and matching of local descriptors in gray-scale images. This method extracts local descriptors invariant to the scale of the object as seen in 2.1.2.1 and was named for this reason SIFT [Lowe 99]. A more accurate description of the method as well as some improvements are presented in a second version in 2004 [Lowe 04].

Section 2.1.2.1 describes how keypoints are extracted in the SIFT algorithm. This section now presents how the descriptors are computed. It is worth noting that both components can be used independently and still achieve good results as reported in a study from Mikolajczyk and Schmid [Mikolajczyk 05].

While SIFT keypoints are already designed to be scale and orientation invariant, the SIFT descriptor is created to provide a higher discriminative power by adding invariance to changes of illumination and 3D points of view. A descriptor is computed from a patch of the image extracted around a keypoint location at the corresponding scale parameter σ. First the local coordinate system is modified to provide rotation invariance. This is achieved by rotating the region by the keypoint orientation parameter but in the opposite direction. Then a patch is extracted at the keypoint location as a 16×16 pixels region subdivided in 4×4 zones of 4×4 pixels each. On each zone is computed an histogram of gradients of 8 bins equally separated. In each pixel the gradient and amplitude are computed. The gradient gives the interval to increment in the histogram which is done by a double weighting: by the amplitude and a gaussian kernel centered on the keypoint, with a spread of 1.5σ (see Figure 2.8). The 16 histograms are then concatenated and normalized. In order to reduce the descriptor's sensibility to changes of illuminations, the values are thresholded at a maximum of 0.2 then the histogram is normalized again to finally create a descriptor of dimension $8 \times 16 = 128$.

HOG and HOF Dalal et al. introduced Histogram of Oriented Gradients (HOG) in [Dalal 05]. The neighborhood of a keypoint is divided into "cells" (small spatial regions) in which a one dimensional histogram of gradient orientation is computed. All cells histograms are then aggregated to create a main histogram of gradient descriptor which is made invariant to illumination changes thanks to a contrast normalization. The same scientist proposed one year later in [Dalal 06] the Histogram of Optical Flow (HOF) descriptor for videos. The method employed is fairly similar to the HOG one but replacing gradient information by the optical flow. HOG and HOF descriptors are often combined together since the work of Laptev et al. in [Laptev 08] where they build a spatio-temporal bag of features with these descriptors for action classification in movies, significantly outperforming state-of-the recognition performances.

The Speeded Up Robust Features (SURF) descriptor Due to their success, local descriptors have become more widespread and researchers have tried to come up with ways to increase their computation efficiency [Bay 06, Cornelis 08, Bay 08]. In 2006 Bay et al. propose a new method to extract and compute local descriptors named SURF (Speeded-Up Robust Features) [Bay 06], which is designed as a computationally efficient alternative to SIFT. This section is dedicated to provide a brief overview of this descriptor as it is the one selected throughout this thesis contributions. The first step of the process is to extract interest points with a Fast-Hessian detector. This detector, similarly to SIFT, is designed to be scale invariant but is based on an approximation of a gaussian filter allowing to greatly reduce the computation times. Bay et al. approximate second order gaussian derivatives by simpler filters

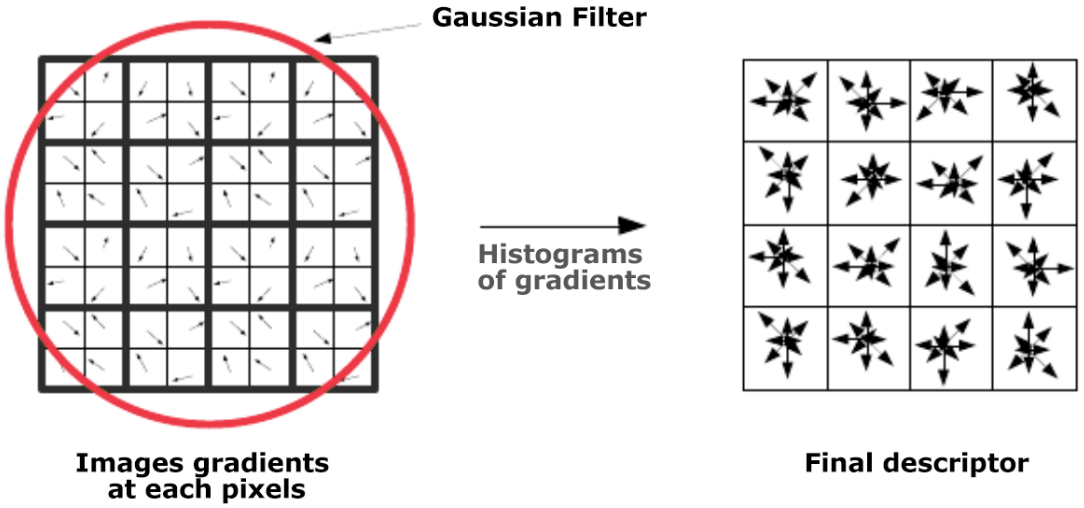


Figure 2.8 – Illustration of the Gaussian weighting of gradients in the SIFT descriptor to create 8 histograms of gradients. Image derived from [Lowe 04]

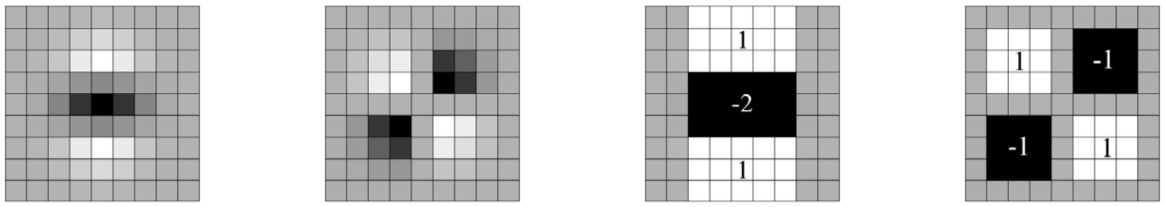


Figure 2.9 – Left to right: the (discretised and cropped) Gaussian second order partial derivatives in y-direction and xy-direction, Proposed approximations using box filters in [Bay 06]. The grey regions are equal to zero. Image from [Bay 06]

(see Figure 2.9). The second step concerns the orientation invariance. The authors use Haar wavelets on the integral image [Viola 04] in order to further reduce the computation time. These wavelets allow to compute the first derivatives of the image around the keypoint and study the distribution of horizontal and vertical gradients from which cant be extracted the main orientation of the keypoint. This step can be seen on figure 2.10.

On the figure the circle represents the region of interest around the keypoint with a radius of 6σ where σ is the scale parameter associated to the keypoint. The angle of the main orientation is derived from the search of the highest distribution of wavelet responses in a $\pi/3$ wide angle zone. The descriptor itself is based on the sum of the horizontal and vertical wavelet responses as well as their norm. Figure 2.11 illustrates the descriptor computation process. The description zone is divided in 16 regions each of them sub-sampled in 25 sub-regions. Wavelet responses in each region are analyzed to construct the final descriptor based on the summary statistics $\sum dx$, $\sum |dx|$, $\sum dy$, and $\sum |dy|$ resulting in a 64 dimensions descriptor.

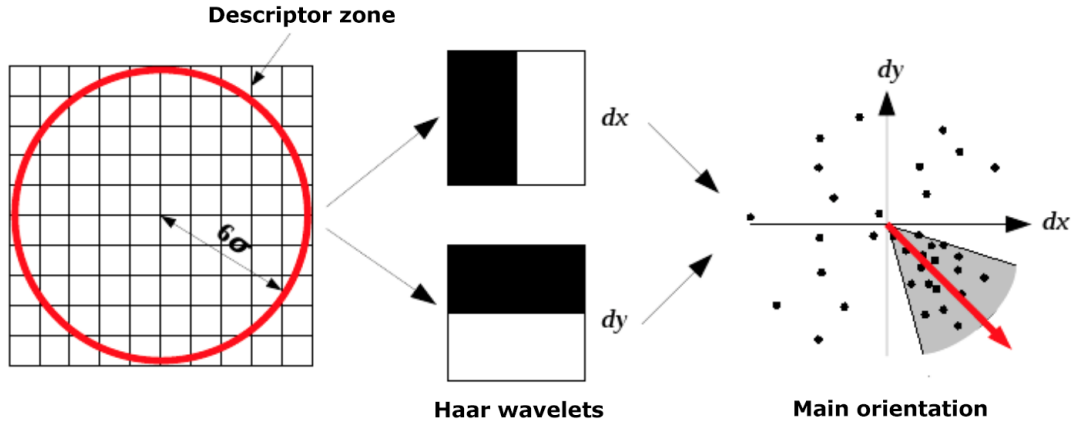


Figure 2.10 – Finding of the main orientation in the SURF descriptor computation. Image derived from [Grand-Brochier 11]

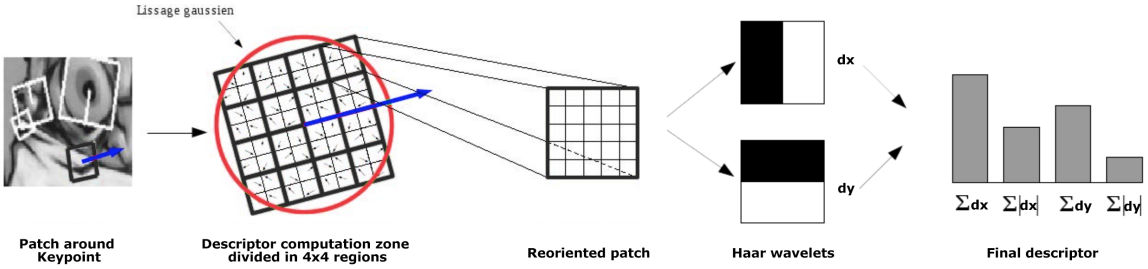


Figure 2.11 – Illustration of the computation of the SURF descriptor. Image derived from [Grand-Brochier 11]

2.2 Matching low-level features

Now that we presented how to extract low level features in images, the next step is to find methods to recognize those same features in different images. This is called the *matching step* and is used in a lot of application for instance recognition. In this section we both discuss and present some example methods of matching and an overview of the applications.

2.2.1 Matching local features

As discussed before in 2.1.1, the advantages of using local features compared to global features are various, that is why this section focuses only on matching local features. Since the properties used for extracting local features are invariant to most real-world transforms, a common matching technique is to describe the model object by a constellation of its local features in the training stage and to search the same spatial arrangement of features in the test image during the detection stage. We distinguish in the following between two different ways of verifying the geometric consistency of a constellation: namely, rigid and non-rigid techniques.

2.2.1.1 Rigid matching

Methods that rely on a rigid transform (e.g. a projective transform) to constrain the local feature positions can be classified into two categories: RANSAC-based methods and Hough-based methods.

Hough-based matching The Hough transform was first patented in 1962 and later adapted for the computer vision community by Duda and Hart [Duda 72]. From all specific object detection methods using the Hough transform, the SIFT algorithm from Lowe [Lowe 04] is probably the most famous and popular. In this context the Hough Transform is used to cluster reliable model hypotheses to search for features that agree upon a particular model pose. When the same object model appears several times for different features it creates a cluster. When clusters of features are found to vote for the same pose of an object, the probability of the interpretation being correct is much higher than for any single feature.

During the training stage, multiple views of the same object are combined in order to compute a set of characteristic views. At the same time, SIFT features belonging to the model views are indexed in a k-d tree in order to enable a fast pairwise matching between models and SIFT features (this technique is scalable to a large number of model objects and thus has been replicated in many other works, e.g. see [Bay 06]).

During the detection phase, the SIFT descriptors extracted from the test image are matched to their corresponding models using the k-d tree. Then, the Hough transform is performed: each matched descriptor votes for its corresponding models in the parameter space. An entry in a hash table is created predicting the model location, orientation, and scale from the match hypothesis. The hash table is searched to identify all clusters of at least 3 corresponding model votes in the parameter space, and the bins are sorted into decreasing

order of size. Finally, each identified cluster is subject to a verification procedure in which a linear least squares solution is sought for the parameters of the affine transformation relating the model keypoints to the currently tested keypoints. At least 3 matches are needed to provide a solution.

The drawback of such an approach is that it does not take into account the real 3D shape and the 3D transformations of the object and therefore is unable to recover its precise spatial pose. Moreover, Moreels and Perona [Moreels 08] have shown that the choice of the bin size in the Hough space is problematic (smaller bins cause fewer true positives, while larger bins may lead to missed detections).

RANSAC-based matching The Random Sample Consensus (RANSAC) algorithm was introduced by Fishler and Bolles in 1981 [Fischler 81]. It is possibly the most widely used robust estimator in the field of computer vision. The RANSAC algorithm can be summarized as follows: assuming a noisy set of samples and a given spatial transform, the algorithm iteratively picks a small number of input samples and estimate the transform parameters of the associated fitting problem. Then, a score is given to this trial to measure its quality, usually by counting the number of inliers, i.e. the number of other samples that comply with this parametrization. Finally, the transform parameters corresponding to the best trial are returned. Because RANSAC relies on a succession of random trials, it is not guaranteed to find the optimal solution. A probabilistic formula is used in practice to determine the number of iterations necessary to output the optimal solution with some confidence. Numerous papers related to the matching of specific objects or even whole scenes, like short and wide baseline stereo matching [Chetverikov 02, Matas 02], motion segmentation [Torr 95] and of course specific object detection [Lepetit 05, Rothganger 06] have used RANSAC coupled with keypoints as robust estimator. Lepetit et al. [Lepetit 05], for instance, have presented a real-time system based on randomized trees for keypoint matching. Their solution is notably robust against changes in view point and illumination. In a different fashion, Rothganger et al. [Rothganger 06] have considered affine invariant keypoints to recover more efficiently the object pose from the matched feature patches. Even if those methods give good results, a common drawback is that the 3D shape of the model objects has to be learned beforehand.

2.2.1.2 Non-Rigid matching

One drawback of rigid matching techniques is that they cannot handle distortions like what happens to a bent magazine or to a moving person. Non-rigid matching, on the contrary, assumes that the model object can be decomposed in a set of different independent parts that can move on their own (with some limits, of course). This strategy has been shown to give more flexibility to the model [Ferrari 06] and to increase performances thanks to the fact that distant features are disconnected [Chum 09]. The matching cost is however often superior compared to the case of a rigid matching, but this is expected as the number of parameters that govern a non-rigid transform is by far superior to the number of parameters for a rigid transform. Non-rigid matching can be roughly categorized into two kinds of techniques: those relying on graph matching and those denoted as part-based models A

large amount of studies have tackled the recognition problem using graph matching, for instance applied to the detection of faces [Wiskott 97], indoor objects [Gold 96] or mechanical parts [Kim 91, Chevalier 07]. The main drawback of this kind of approaches however lies in the computational power needed to match two graphs. Apart from graph matching, a closely related field is the class of part-based object recognition methods. Although the term "parts" may refer to semantic parts (especially for class object recognition methods, see next section 2.3), we restrict here to the case of specific objects. In this context parts thus only mean local patches of the object surface, most often derived from sparse feature detectors. Part-based models for specific object recognition address the problem in a similar fashion than graph matching (i.e. decomposing objects in parts loosely connected) but use different techniques to solve the part assignment problem [Ferrari 06, Detry 08, Holzer 09]

2.2.2 Matching for Instance recognition and other applications

Matching local features is still too "object-specific" to be applied to the task at end in this thesis, which is the recognition of object categories. However this specificity is useful in several other applications. In the following, we present some applications where the specific object recognition techniques presented before are used in practice. The purpose of this overview is to give the reader a feeling of the range of possibilities, but it should by no means be thought of as an exclusive list.

The first application we can mention is of course specific object recognition which has been made possible but also robust and efficient by the introduction of local scale and rotation invariant features such as SIFT [Lowe 04]. The approach described before in Section 2.2.1.1 based on the Generalized Hough Transform is one of the most popular example of these approaches [Lowe 99, Lowe 04]. Another motivation for the development of such affine invariant local features is their use for wide-baseline stereo matching. Here the aim is to find correspondences between a model view and a set of other views. Extending this idea, matching local features has allowed to develop robust and efficient panoramas creation techniques [Brown 07]. Efficient similarity search algorithms such as kd-trees [Friedman 77] or Hashing-based algorithms [Indyk 98, Gionis 99] allow to apply recognition paradigms to very large data sets. This is called *large scale image retrieval*. One particularly useful application of large scale image retrieval nowadays is the visual search from smartphones. The goal is to make it possible for users to perform a search of something specific based on a picture or a photo taken by the smartphone and get relevant information about it. Among the first works to propose a practical implementation of large scale mobile visual search was proposed in 2006 by Nister and Stewenius [Nistér 06]. It was based on local features and a Vocabulary Tree indexing scheme. Their approach can recognize a picture of a CD cover in a database of up to 50,000 different covers in less than a second (with a single laptop with 8GB of RAM). Nowadays several different commercial services offer large-scale visual search applications such as Google goggles (www.google.com/mobile/goggles/), kooaba Visual Search (http://www.kooaba.com/), or Amazon Remembers. Since most approaches are based on local features, previously described in section 2.1.2, they are particularly adapted to recognizing textured, planar objects, such as book/CD/DVD covers, movie posters, or

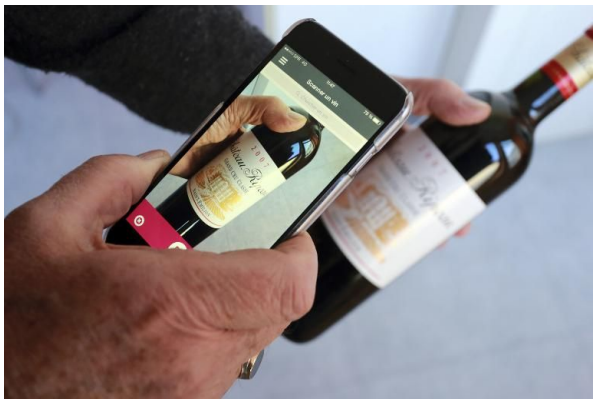


Figure 2.12 – Example of a commercial application: "Winewoo" smartphone application developed in Bordeaux about wine bottles labels recognition.

even wine bottle labels (see Figure 2.12).

2.3 Class object recognition

As we saw previously, simple features such as keypoints are enough for specific object recognition. Although using the same features as well for classes of objects could appear to be a good idea, it is not that simple. The main problem lies in the fact that simple features are often too specific, enabling little generalization regarding the larger intra-class variations occurring for classes (see fig 2.1b). To overcome this issue, category recognition paradigms have to aggregate the extracted low-level features into higher level models designed to be more invariant to intra-class variations.

In this section we start by presenting two well-known kinds of representations for class object recognition, namely the BoVW model and the Deformable Parts models. In a second part this section will introduce several common detection schemes for classification.

2.3.1 Different representations for object classes

There is a lot of work about finding ways to represent images by capturing and summarizing relevant visual cues in order to recognize not instances but whole categories of objects. In this section we briefly overview some candidate approaches. One noticeable fact is that, despite the variety of possible representations for object categories, most of them fall in one of these two categories: *window-based models* or *part-based models*. Window-based models describe the whole data inside a region of interest while part-based models define the data contained in local parts taking into account the geometric structure connecting them. In this part, we start by presenting some of the main approaches for window-based models based on the Bag of Visual Words approach and follow by introducing basics of part-based models.

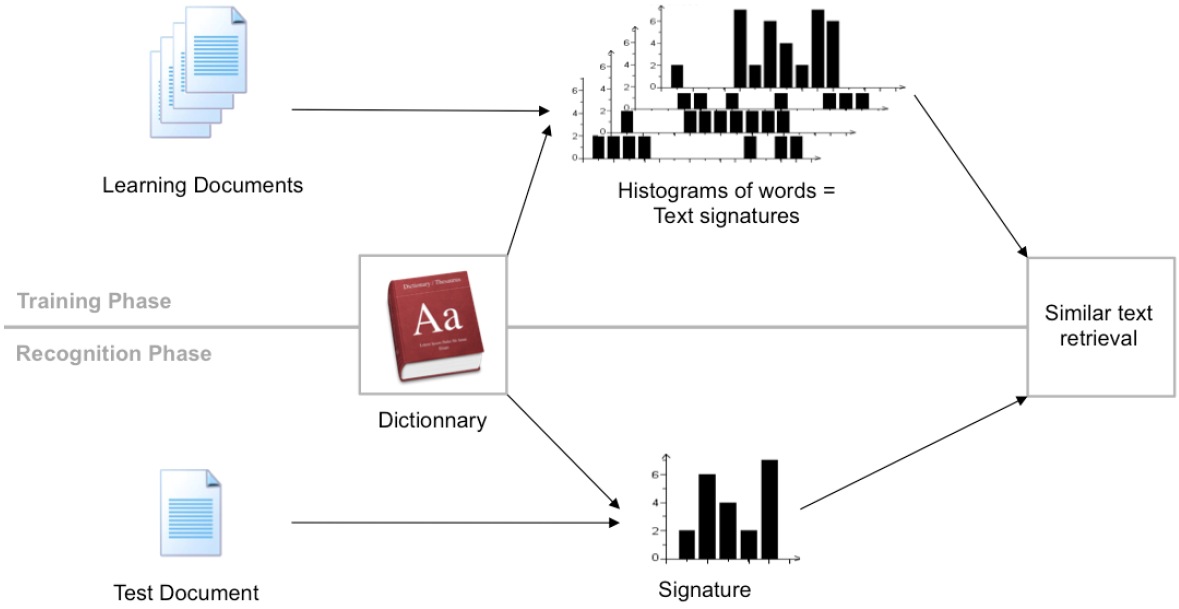


Figure 2.13 – Illustration of the BoW concept for text retrieval.

2.3.1.1 Bag of Words models

The state-of-the-art in image or object categorization and recognition has been highly influenced by the paper [Sivic 03] published by Sivic and Zisserman. In their paper, they proposed to apply many techniques from text retrieval applications to the visual content retrieval task. The BoW framework will first be presented for the application to text documents. Then, the main steps for its application to images will be reviewed.

Concept of Bag-of-Words for text documents ‘

In text retrieval [Lewis 98], documents are parsed in words. Each word is represented by its stem, for example the stem «walk» stands for the possible variations «walking» or «walks». Then a stop list is used to reject the most common words such as «the» and «an» since they are not discriminant. A unique identifier v_i is associated to each stem. Each document d is represented by a vector W giving the frequency of occurrence of the words the document contains: $W_d = (t_{v_1}, ..., t_{v_i}, ..., t_{v_k})$.

These values may be weighted, for example by the Term Frequency-Inverse Document Frequency (tf-idf) weighting [Salton 86]. Each component of the vector representing the document is the weighted word frequency computed as the product of two terms as in eq.(2.9): the *word frequency* $\frac{n_{id}}{n_d}$ and the log *inverse document frequency* $log\ \frac{N}{n_i}$, where n_{id} is the number of occurrences of word v_i in document d, n_d is the total number of words in the document, n_i is the number of occurrences of term v_i in the whole database and N is the number of documents in the whole database. The word frequency term weights words occurring often

in a document while the inverse document frequency term down weights words that appear often in the database and are therefore less discriminative.

$$t_{v_i} = \frac{n_{iI}}{n_I} log \frac{N}{n_i} \tag{2.9}$$

Another interesting technique is the use of inverted file which enables fast retrieval. An inverted file has an entry for each word in the corpus followed by a list of documents in which the word occurs. Finally, a text is retrieved by computing its vector of word frequencies and returning the documents with the closest vectors.

The BoVW framework has been built using he same idea as BoW. It has four main stages: building a visual dictionary, quantifying the features, choosing an image representation using the dictionary and comparing images according to this representation. These steps are explained in the following paragraphs.

Extending Bag-of-Words to images: building a Visual dictionary When using images which are only composed of pixel values in a color space, it is necessary to define an equivalent to words in the text context. The images are represented by a set of features describing the content of some regions of interest extracted from the image. Local features such as SIFT and SURF introduced in section 2.1.2.2 are relevant and widely used for image representation. Local features computed over the same object or part of an object contained in different images have many variations due to different factors such as different points of view, intra-class variations... They can be seen as hand written words of variations of a stem. According to this analogy, it is necessary to create a set of «visual words» that we can call a «visual dictionary» and denote it by V of K visual words. Generally, a set of randomly selected features is used to build a visual dictionary by clustering. Similarly to the method in text domain, the most common and rare words can be deleted from the dictionary to enhance the performance.

In the initial BoVW framework proposed by Sivic and Zisserman [Sivic 03], the visual dictionary was built by a k-means clustering [MacQueen 67]. This method has been widely used since [Csurka 04, Winn 05].

Feature quantization In order to build a robust representation of an image that has some desirable properties such as compactness, sparseness (i.e. most components are 0) or statistical independence, the feature vectors are quantized according to the visual dictionary V. Usually the quantization step consists in assigning each feature f_i of an image to its closest word v_j in the dictionary V, with $j = 1..K$. This process can be referred to as the «coding step». However different coding operators have been proposed in the literature. In [Boureau 10], Boureau et al. have studied the influence of three of the most widely used coding operators: hard quantization, soft quantization and sparse coding. In the following paragraphs we will review the principles of these processes.

Hard quantization Hard quantization is the classical formulation of the bag-of-words framework [Sivic 03]. The coding operator q minimizes the distance to a code book, i.e. each feature f_i is assigned to the closest codeword in the dictionary, which is usually build by an unsupervised algorithm such as K-means. Let v_j denote the j-th codeword. This process can be formalized as:

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } j = \underset{j}{argmin} \parallel f_i - v_j \parallel_2^2 \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

Soft quantization From the previous definition it is clear that there is a strong quantization by assigning a continuous feature to a single representative. This drawback has been studied by Gemert et al. in [van Gemert 10]. They explore soft quantization techniques and evaluate the influence on classification performances when using low to high dimensional features or small to very large vocabulary. The idea of soft quantization is to tackle the ambiguity of a visual word that hard quantization simply ignores. The drawbacks of hard quantization are twofold: (i) when a data sample is close to several codewords, only the closest is considered and (ii) a codeword is assigned to the closest codeword no matter how far it can be. The first aspect is referred to as *word uncertainty* and the second as *word plausibility*. Instead of using histograms to estimate the probability density function the authors proposed to use a kernel density estimation. Defining a Gaussian-shaped kernel $K_\sigma(x)$ in (2.11), three models are studied in this paper. The *kernel codebook* KCB defined in (2.12) will weight each word by the average kernel density estimation for each data sample. The *codeword uncertainty* UNC defined in (2.13) normalizes the amount of probability mass to a total of 1 which is distributed over all relevant codewords. Finally, the *codeword plausibility* PLA defined in (2.14) will give a higher weight to more relevant data samples but is not able to select multiple codeword candidates. The results on a classification task obtained on the data sets Scene-15, Caltech-101, Caltech-256 and Pascal VOC 2007/2008 shows that the *codeword uncertainty* UNC outperforms all other methods using either low or high dimensional feature vectors or small or very large visual vocabulary.

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\, exp(-\frac{1}{2}\frac{x^2}{\sigma^2}) \tag{2.11}$$

$$\alpha_{i,j}^{KCB} = K_\sigma(\, D\,(v_j,\, f_i)) \tag{2.12}$$

$$\alpha_{i,j}^{UNC} = \frac{K_\sigma(\, D\,(v_j,\, f_i))}{\sum_{k=1}^{|V|} K_\sigma(\, D\,(v_k,\, f_i))} \tag{2.13}$$

$$\alpha_{i,j}^{PLA} = \begin{cases} K_\sigma(\, D\,(v_j,\, f_i)) & \text{if } v_j = \underset{v_j \in V}{argmin}(\, D(v_j,\, f_i)) \\ 0 & \text{otherwise,} \end{cases} \tag{2.14}$$

Sparse Coding Sparse coding [Olshausen 97] uses a linear combination of a small number of codewords to approximate the feature f_i. These codewords are represented by a dictionary $V = (v_1, \ldots, v_k)$ in matrix form $V \in \mathbb{R}_{d \times K}$ where d is the dimension of the feature space. The linear weights correspond to the vector $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,k})^T \in \mathbb{R}^K$. Yang et al. [Yang 09] have obtained state-of-the-art results by using sparse coding and max pooling. We will here only briefly define the sparse coding process as an overview of sparse coding is beyond the scope of this manuscript.

The sparse coding cost function is defined for the vocabulary V and coding coefficients α_i as:

$$L(\alpha_i, V) = \|f_i - V\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 = \left\| f_i - \sum_k \alpha_{i,k} v_k \right\|_2^2 + \lambda \|\alpha_i\|_1 \tag{2.15}$$

where $\|\alpha\|_1$ denotes the L_1 norm of α, λ is a parameter that controls the sparsity. The dictionary V is trained by minimizing the average of $L(\alpha_i, V)$ over all samples, alternatively over V and the α_i. Yang et al. and Boureau et al. [Boureau 10] have shown that sparse coding outperforms both hard quantization and soft quantization in several data sets such as Scene-15, Caltech-101 and Catlech-256.

Image representation According to the visual dictionary of K words, each image I of the dataset can now be represented by a K-vector of visual word frequencies W_I. Usually, the vector is normalized by the number of features within the image. Therefore, W_I is a normalized histogram representing the distribution of visual words for the image I.

Limitations and improvements The BoVW framework was clearly a breakthrough in the domain of image recognition or retrieval. However, this framework had some limitations that have been discussed and challenged since the paper of Sivic and Zisserman [Sivic 03].

One of them is that the BoVW model based on interest points usually ignores the spatial (or geometric) information, i.e. the spatial or spatio-temporal relationships between the interest points. These relationships contain many important complementary clues for recognition. In [Grauman 05], Grauman and Darell propose a pyramid match kernel, which maps unordered features to multi-resolution histograms and implicitly finds correspondances on the finest resolution. Labzebnik and Schmid [Lazebnik 06] partition the image into increasingly fine sub-regions and compute a BoVW model for each sub-region. Cao et al. [Cao 10] encode geometric information of objects into ordered BoVW models. A histogram transformation is applied to get a spatial bag-of-features, which is tolerant to variations in translation, rotation, and scale. Ren et al. [Ren 14] present a Bag-of-Bag approach derived from [Lazebnik 06] by proposing an irregular partitions (subgraphs) of images are further built via Normalized Cuts [Shi 00] instead of partitioning into increasingly fine grids.

In [Perronnin 07], Perronnin et al propose to apply Fisher kernels to image categorization. The Fisher kernel idea is to characterize a signal with a gradient vector derived from a probability density function (PDF) which models the generation process of the signal. This representation can then be used as input to a discriminative classifier. For the problem of

image categorization they propose to use images as input signals and to use as a generative model a GMM which approximates the distribution of low-level features in images. Their approach became very famous since it showed excellent performances on various challenging databases (and notably the PASCAL VOC 2006 database) while being computationally efficient. Later in [Avila 13], the authors proposed the BossaNova approach for improving the pooling step of BoVW frameworks. When pooling, there is a compromise between the invariance obtained and the ambiguities introduced. If different concepts represented in the image end up activating sets of codewords that overlap too much, ambiguities can arise and the following step of classification will have difficulty in separating those concepts. One way to mitigate that problem is to preserve more information about the encoded descriptors during the pooling step. Instead of a simple sum of the activations, like in the classical BoVW, more detailed information can be kept. In BossaNova, the authors propose to estimate the distribution of the descriptors around each codeword. They consider a non-parametric estimation of the descriptors distribution, by computing a histogram of distances between the descriptors found in the image and each codebook element

2.3.1.2 Overview of Deformable Part-Based Models (DPMs) for class object recognition

In DPMs an object is represented as a set of meaningful parts and their associated spatial structure. Compared to window-based models such as BoVW models, part-based models, are designed to better express the structural properties within objects and, to a certain amount, are able to cope for their changes. The first DPMs have been proposed in [Fischler 73] by Fischler and Elschlager in 1973. They proposed the first *pictural structure* model, a type of DPM where the structure of the parts modeling an object is captured by a set of "springs" connections (Gaussian distribution between pairs of parts). At the core of their proposal is an energy-based function based on the sum of the mismatch cost of each part and the deformation cost between pairs of parts. The best configuration of an object is the one which minimises this energy function.

Since then much work has been done regarding these models [Fergus 03, Mikolajczyk 06, Zhu 10]. Recently the model from Felzenszwalb et al. [Felzenszwalb 10] has achieved state-of-the-art results for object localization on many categories in the recently PASCAL VOC challenges. Felzenszwalb et al. [Felzenszwalb 10] attempt to learn deformable part based models in a discriminative way. In this approach one novelty is that they introduce latent variables to model the parts so that they can be learned automatically and efficiently without previous user annotations of parts in the training set. Since this model has become a reference and has been efficiently used for active object recognition in egocentric videos [Pirsiavash 12], we briefly review its definition in the following.

In this model the DPM is composed of a root filter P_0 and a set of part filters $P_i, i \in (1, ..., n)$. Each part filter also defines a two-dimensional vector $a_i = \langle a_{ix}, a_{iy} \rangle$ specifying the anchor position relative to the root, and a four quadratic deformation coefficient d_i taking into account the cost of the part relative to its anchor position.

The hypothesis score of a window is the sum of the scores form the root filter and part filters, minus the deformation cost of each part location regarding its anchor position (see

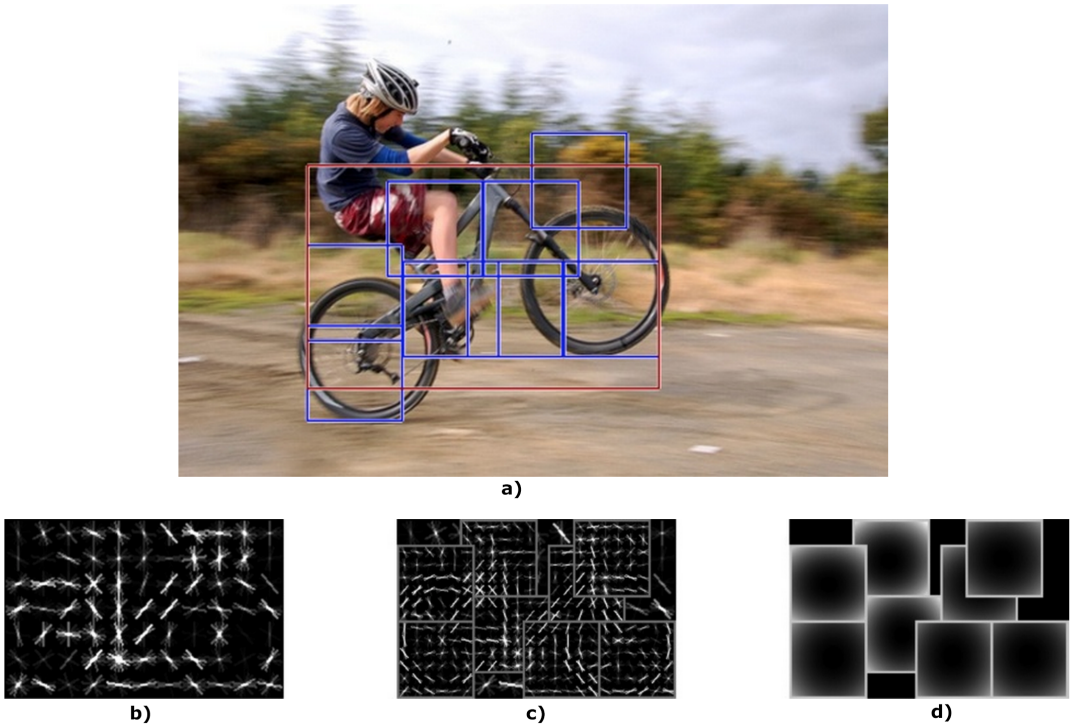


Figure 2.14 – Illustration of a DPM for the category "Bike". a) source image, b) root filter, c) part filters, d) deformation cost

Figure 2.14), and plus a bias b, as described follows:

$$f(P_0, ..., P_n) = \sum_{i=0}^{n} P_i \cdot \phi(x, z_i) - \sum_{i=1}^{n} d_i \cdot \psi(z_i) + b \tag{2.16}$$

where z_i is the inferred part location, and $\phi(x, z_i)$ are local dense histograms of oriented gradients features extracted from z_i, $\psi(z_i)$ is the displacement of the inferred position z_i with regard to the anchor position a_i by:

$$\psi(z_i) = [(z_{ix} - a_{ix})^2, (z_{ix} - a_{ix}), (z_{iy} - a_{iy})^2, (z_{iy} - a_{iy})] \tag{2.17}$$

The part locations, treated as latent variables, are inferred in the learning and test stage to maximize the score. Equation 2.16 can be rewritten as:

$$f(P_0, ..., P_n) = \max_{z \in Z(x)} (\sum_{i=0}^{n} P_i \cdot \phi(x, z_i) - \sum_{i=1}^{n} d_i \cdot \psi(z_i) + b) \tag{2.18}$$

where $Z(x)$ is a collection of all possible positions for the parts. The objective function to minimize is the hinge loss, which is expressed by:

$$L(\beta) = \frac{1}{2}\|\beta\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i f(x_i)) \tag{2.19}$$

where β is the set of learning parameters containing $(P_0, P1, ..., P_n, d, b)$ x_i and y_i are the training example and its ground truth, $f(x_i)$ is its score, C controls the regularization term, N is the training instance size. The equation 2.19 is then minimized using latent SVM. The training stage learns part filters and deformation costs to infer part locations in the test phase. In this framework, anchor positions are automatically set to locations where the root filter gives high responses, this might not lead to optimal results but no human part annotation is needed.

DPMs have received many improvements, for instance by sharing parts between object classes [Ott 11], or by defining the components for each class with visual similarity instead of height-width ratio with the ground-truth bounding boxes [Divvala 12]. DPMs have also been extended to spatio-temporal data for activity recognition [Tian 13]. An implementation from Dean et al. [Dean 13] have greatly improved the efficiency of DPMs using locality sensitive hashing instead of convolutions (enabling the detection of 100,000 objects on a single machine). Recently, in [Girshick 14], Girshick et al. show how DPMs can be formulated as a Convolutional Neural Networks (CNN) by mapping each step of the DPM inference algorithm to an equivalent CNN layer. Their so-called *DeepPyramid DPM* approach significantly outperforms DPMs based on histograms of oriented gradients features on the on PASCAL VOC object detection dataset at very efficient computational rates.

2.3.2 Detection schemes

Now that the reference class-object representation have been presented (lest section 2.3.1), the detections processes can be outlined.

Before starting this section we want to refer the reader to another strategy which takes care of image feature extraction, representation and classification in the form of deep learning networks. Recently, deep learning models have attracted a lot of attention due to the large success of deep convolutional nets in the Large Scale Visual Recognition Challenges [Szegedy 15]. The results reveal that deep learning significantly outperforms competitors using BoVW models [Theriault 11, Theriault 13]. However, although this trend is unquestionable for this large-scale context (1 million training examples), the feasibility of reaching state-of-the art performances in other datasets with fewer training examples, e.g. the case at end or even the well-known PASCAL VOC, remains unclear. It is one of the new perspectives we would like to explore in the future as stated later in 7.2.2.

We also outline the fact that detection strategies for image-based and part-based models differ. Hence we inform the reader that there is an extensive amount of work on detection schemes for deformable part models (introduces in section 2.3.1.2) but, as previously mentioned, we did not use this kind of object representation in our contributions apart for comparison purposes. For the reason of clarity and space, we prefer not to develop the different detection schemes related to DPMs in this section. Instead we will focus this section on a deeper description of detection schemes we used or in direct relationship with our contributions.

2.3.2.1 General classification schemes

Once an image is represented by a window-based model such as BoVW, the basic approach for category detection is an image classification problem. Assuming that classifiers [Friedman 77, Cortes 95] have been trained on the window models for each class it is possible for a test image to be classified on its whole as one of these trained models with a certain decision value. Using a classification step for detection has clear advantages. It is can be implemented in an easy fashion and benefit from the potential of very powerful and sophisticated machine learning algorithms to learn complex patterns. In fact the combination of well-performing local features with learning algorithms has proved to be very effective for several frameworks such as faces or pedestrian recognition [Viola 01, Dalal 05].

2.3.2.2 Window-based schemes

Objects might be cluttered, are not necessarily unique, centered, aligned and cropped in images, especially in real-life situations. To face these problems one possible improvement to the general classification framework previously defined is to insert a sliding-window search into the pipeline. The classifier will then test all possible sub-window of new image and determine in each one the object presence/absence. On top of testing sub-windows of different sizes it is possible to to run a multi-scale search, by resampling the test image and make a scale-pyramid. Since generally the classifier detects multiple times an object around its location for all considered nearby windows, a non-maximum suppression step is usually employed as a post-processing step to reduce the detections to a unique window.

An exhaustive search of all possible windows is obviously very computationally demanding and several approaches have been developed to optimize the search. The idea is to prioritize the search by discarding windows that seem unlikely to contain an object. A sequence of tests are performed on the windows which are designed to be sequentially more computationally expensive if the window does not get discarded along the process [Fleuret 01, Viola 04, Lampert 08]. With this methods, clear negatives are discarded early on and finer tests and classification are performed on windows most likely to possess objects. For instance in the well-known method from Viola and Jones [Viola 04], a cascade structure is implemented to detect more features and produce lower false-positives at each finer detection step. In 2008, Lampert et al. [Lampert 08] propose a branch-and-bound search, dramatically reducing the time for sliding-window search in an image. They achieve this by computing a bounding function for every window easily computable with certain types of simple SVM kernels [Cortes 95].

Another limitation with this method is the rectangular axis-aligned window used for the scanning. Indeed objects are generally far from box-shaped and performing the detection on rectangular windows might introduce corrupted information. One could also agree that considering only windows independently from the rest might be a handicap since it discards possible contextual information.

For these reasons as well as the still heavy computational burden of sliding window-methods, other strategies have been introduced. The next sections will present segmentation-based and saliency-based stategies conceived to address these issues.

2.3.2.3 Segmentation-based schemes

In order to overcome the drawbacks of sliding windows methods previously presented, one of the proposed solution is to add a segmentation pre-processing step to the recognition pipeline. This idea is to over-segment an image as a first step, in order to get thousands of object segments candidates. Then bounding boxes are created around the regions and fed to the previously trained classifiers. One of the recent approach developing this idea is the *Segmentation as Selective Search (SS)* [van de Sande 11] approach. After the over-segmentation step, a hierarchical segmentation tree is constructed by grouping similar regions with a greedy algorithm. Then the classification is performed on bounding boxes built around these regions across the tree. This method allows to use more computationally costly features for each region and showed an improvement over the state-of-the-art for 8 out of 20 classes on Pascal VOC 2007 dataset.

In [Uijlings 13], Uijlings et al. apply this SS approach to extract a small set of small regions they name *regionlets*. They aim to capture the spatial layout of the object by organizing together several groups containing regionlets. They aggregate these features into a vector designed to tolerate small deformations and then classify bounding box candidates using a previously trained cascaded boosting classifier. This approach is used as a comparison method in section 5.3.4. Recent work by Girshick et al. [Girshick 13] obtained a large improvement in mAP on PASCAL VOC 2011/2012 datasets. They also applied the SS [van de Sande 11] approach to generate a set of candidate bounding boxes, then crop the

content of each bounding box and feed it to a CNN to extract and classify features.

2.3.2.4 Saliency-based schemes

The BoVW model [Sivic 03] is still one of the most prevalent approaches due to its simplicity. However, its performance is greatly limited in case of occlusions or small objects in cluttered backgrounds. In contrast, sliding window methods have turned out to be more robust against these problems. They perform a window-based scanning process that searches for objects in several locations and scales in the image, thus addressing both the detection and accurate localization of objects even when they are small. Nevertheless, these methods still suffer from several drawbacks as stated before: a) although efficient implementations exist, the computational complexity due to the computation of features within each candidate window and the evaluation of the objective function cannot be neglected; b) they require a strong human effort to manually annotate bounding boxes in the training data; c) an exhaustive scanning might cause more false detections; or d) unless explicitly incorporated, context information around the object is usually discarded in this family of methods.

Alternatively, modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered *of interest* or *salient*. Due to the use of saliency maps, the search for objects in images is more focused, thus improving the recognition performance and additionally reducing the computational burden. Even more, saliency methods can be naturally applied to both BoVW [Ren 10] and sliding window approaches [Alexe 12, Uijlings 13].

In general, previous approaches tackling this problem can be broadly divided into three categories: methods using *binary segmentation masks*, *saliency-based pooling*, and *saliency-based sampling*.

Traditionally, most works have relied on binary saliency maps, also known as foreground masks, as a way to delimit the particular area of the image to be processed in discussed in part 2.3.2.3. This is the case of [Walther 04], where object matching is improved by filtering out the local descriptors located in non-salient areas, or the more recent proposal [Ren 10], where the authors incorporated foreground masks to the BoVW paradigm by restricting the detection of local features to particular salient areas of the image. A similar approach is followed in [Fathi 11b], where a method for object recognition in egocentric video is proposed that firstly identifies foreground areas in each frame and consequently detects and labels regions associated with the hands and the object being manipulated.

Following the second strategy, the works in [de Carvalho Soares 12, San Biagio 14] substitute these binary masks by a soft-pooling scheme over real-valued saliency maps. In particular, both works build over the BoVW paradigm, and consider the continuous values of a saliency map to weigh the contribution of each visual word. In addition, in [de Carvalho Soares 12] two complementary image signatures are considered: one associated with the foreground, and another modeling the background. These signatures enable foreground and background-based object recognition, or even combined recognition in which both the object of interest and the context are considered. In [Sharma 12], a discriminative approach for pooling visual features is proposed that integrates both the computation

of saliency maps and the learning of SVM-based classifiers within a unified framework. In this case, saliency maps are category-dependent functions that learn the spatial distribution of visual words associated with particular object categories. This approach has been successfully applied to various computer vision tasks, such as action recognition or scene classification.

Concerning the third category of methods, other works have used saliency to perform non-uniform sampling of local features in images, so that more information is gathered on those areas considered as salient. In [Moosmann 06] the authors propose a classification method based on the use of decision trees over randomly sampled square patches of different sizes. To improve this random sampling process, category-specific saliency maps store the most likely locations and scales of positive patches of each class. The works in [Vig 12] and [Mathe 12] also explore the same idea in the BoVW paradigm, so that local descriptors are computed over regions randomly sampled using saliency maps. Finally, [Alexe 12] and [Uijlings 13] are yet other examples of this kind of approach, where saliency maps drive the search process of sliding-window object detectors, thus drastically reducing the number of windows being evaluated.

2.4 Conclusions

In this chapter, we provided a state-of-the-art of visual recognition. We discussed the commonly used feature types and described the ones we considered most relevant for the rest of our work. Then we introduced how to match these features and what kind of real-life applications can emerge from it. Finally we discussed two representations for class object recognition, namely bag-of-visual-words methods and part-based models, and presented different detection schemes. We finished the list of detection schemes by the one we investigated throughout this manuscript, that is to say the saliency-based detection scheme. Chapter 5 provides a systematic study of the application of saliency to the challenging task of active object recognition but before that, we need to introduce what is visual saliency. Therefore the next chapter will be dedicated to providing a state-of-the-art about visual saliency modeling.

Chapter 3

State of the art in Visual Saliency Modelling

Even if visual perception seems like an easy and natural task for people, the brain does not actually possess the computational ability to analyse the vast quantity of visual data entering our eyes every second. Moreover it needs not only to capt this information but also to analyze and translate it in order to extract the knowledge needed to understand the scene. Generally speaking a "scene" comports several objects, in different positions, under different points of view, at different depths, in movement or not. In addition each object's characteristics possess several possible variations (shape, color, size, texture...). To address the problem that processing all this data in real-time is unfeasible, the brain has developed clever mechanisms to reduce the amount of unimportant visual data. Indeed our visual scene perception is built only from the information that grabs our attention. This process of keeping only relevant information is called *visual attention* and is rendered possible by the HVS (see Figure 3.1)

The basis of many attention models dates back to Treisman and Gelade's Feature Integration Theory (FIT) [Treisman 80]. According to the FIT, a visual scene can be decomposed, similarly to the way the brain does, in simple visual artifacts. These artifacts can be more of less salient and are combined by the HVS in order to attract the attention. A few years later, Koch and Ullman [Koch 85] introduced a way to combine these features using a feed-forward model. They introduced a winner-take-all neural network that selects the most salient location and employs an inhibition of return mechanism to allow the focus of attention to shift to the next most salient location. One could argue they also came up with a first definition of saliency map which is a topographically arranged map that represents visual saliency of a corresponding visual scene (see Figure 3.3)

Several systems were then created implementing related models which could process

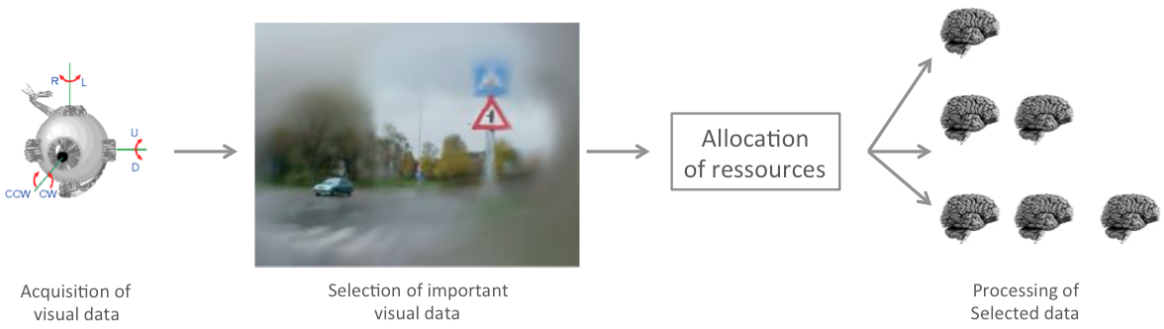


Figure 3.1 – Illustration of HVS automatically selecting regions of importance to process.

(a) (b)

Figure 3.2 – Original image 3.2a and a corresponding grayscale saliency map 3.2b. Figure from [Meur 06]

digital images [Baluja 94, Tsotsos 95]. One of the first complete implementation and verification of the Koch and Ullman model was proposed by Itti et al. [Itti 98] and was applied to synthetic as well as natural scenes. Since then, there has been increasing interest in the field. Computer vision scientists have proposed several models for various applications such as object recognition, quality assessment, image and video compression, color reproduction, robotics, etc... In the State-of-the-art in Visual Attention Modelling [Borji 12a], Borji and Itti present and study the inventory of visual attention characteristics which are considered in visual attention modeling.

While many models have been proposed, it becomes obvious that finding a universal one is not feasible. Indeed even if studies such as the FIT allow to understand which low level information the brain perceive and why some are more important than others, predicting ocular movements still remains a competitive task since visual attention mechanisms rely on both low and high level information. Predicting high level information is very complex, even though some patterns have been identified since many decades. In 1967, Alfred L. Yarbus has shown in Eye Movements and Vision [Yarbus 67] that eye movements are task driven. In his experiment, subjects have been asked to look at pictures. He has noticed that depending on the question he have asked, subjects eye movements have been considerably different. More recently, Michael Land et al. [Land 99] has demonstrated that eye movements anticipate actions when performing IADLs. This means that learning has also an influence on visual attention when a task is accomplished. It would appear that visual attention is the result of a high level reasoning, depending on the task to accomplish, and low level stimulus coming from the visual scene. The main difficulty is to understand how high and low level information are merged. Sometimes, the low level features are more relevant, and sometimes this is the reasoning. Nowadays, this border is still fuzzy and is the subject of many studies.

This section of the manuscript introduces important definitions and characteristics about visual attention modeling, provides a non-exhaustive list of salient features and describes common ways to evaluate how visual attention models are compared with human visual attention.

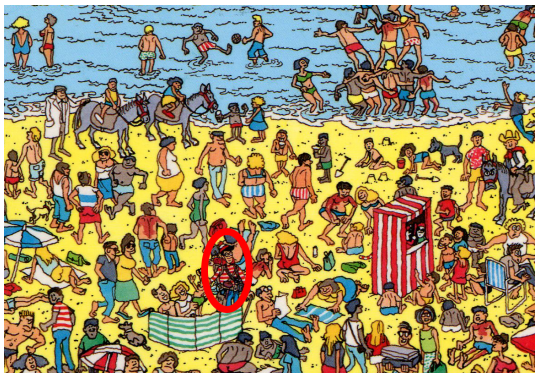
3.1 Visual attention model characteristics

We start by introducing characteristics that will be used later for categorization of attention models. These characteristics have their roots in behavioral and computational studies of attention.

3.1.1 Bottom-up or Top-down

The process of focusing our attention towards a specific region in the visual field is influenced by two factors: one called *bottom-up* and the other *top-down*. Bottom-up saliency refers to exogenous mechanisms guided uniquely by stimuli present in the visual field without any will power from the observer. Therefore bottom-up saliency takes its roots simply from

(a)



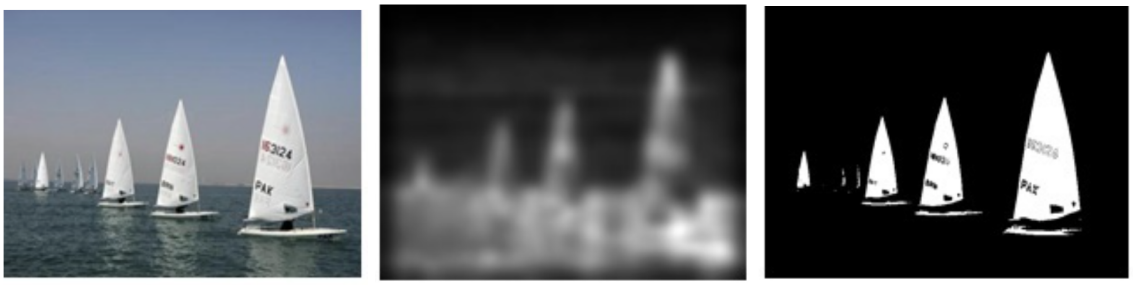
(b)

Figure 3.3 – Illustration of exogeneous (bottom-up) processes in 3.3a which are unconscious, fast, driven by visual properties (here the red computer) and endogeneous (top-down) processes in 3.3b which require a cognitive effort and are slower because semantically driven.

the content of what distinguishes a certain region from its neighborhood based on visual attributes (see Figure 3.3a) Oppositely, top-down saliency refers to endogenous mechanisms based on the will power of the user. These mechanisms are linked to the task being operated but also to the semantic of the stimulus and the own experience of the observer. For example a trained policeman might not look at a crime scene the same way a computer vision scientist would (see Figure 3.3b). Many studies have been conducted in order to characterize the processes of visual attention, whether they are bottom-up [Itti 98, Harel 07, Brouard 09] or top-down [Yarbus 67, Carrasco 11, Pinto 13] and also to find the link between the two [Melloni 12]. They have shown that bottom-up mechanisms are faster and precede the top-down influences, longer for the brain to recognize but lasting longer in time afterwards [Parkhurst 02, Tatler 05]. In the next chapter 4, we present our contributions in both bottom-up and top-down domains for the specific task of active object recognition in egocentric videos.

3.1.2 Saliency and Gaze

While the terms *attention*, *saliency*, and *gaze* are often used interchangeably, each has a more subtle definition that allows their delineation. Attention is a general concept covering all factors that influence selection mechanisms, whether they are scene-driven bottom-up or expectation-driven top-down. Saliency intuitively characterizes some parts of a scene which could be objects or regions, that appear to an observer to stand out relative to their neighboring parts. The term "saliency" is often linked to pixel-based saliency maps which provide a saliency measure for each pixel. Gaze, also referred as scanpath, is a coordinated motion of the eyes and head. Gaze can be built from features in order to predict the fixations order over the visual scene. This is the case for methods proposed by C. Kock et al. [Koch 85], L. Itti et al. [Itti 98], or D. Walther et al. [Walther 06] Similarly, a pixel-based saliency map may be built afterward by using the scanpath.



(a) (b) (c)

Figure 3.4 – Illustration of a) a frame and its corresponding b) fixation prediction (saliency map) and c) salient object detection map. Source: [Borji 14]

3.1.3 Overt or covert attention

Researchers have classified two kinds attention: the hidden attention or *covert* attention and the flagrant attention or *overt* attention. Studies have shown that the attention can move independently than ocular movements. [Rayner 98, Henderson]. The overt attention is characterized by the fact that the eyes are moving towards a particular region where the observer is paying attention. Oppositely covert attention is defined as the process focusing on a specific region of the visual scene without ocular movements towards this scene. Studies also show that the perspective of of future ocular movements towards a region originates form a shift of the attention towards this region that precedes eye movements [Hoffman 95, Rayner 98]. These two kinds of attention have been highlighted during experimental studies in laboratory conditions. For certain specific tasks, where complex stimuli are involved such as natural scenes or reading scenes, some studies show that it is more natural to move eyes where the attention is focalized rather than focalize attention without eye movements [Torralba 06]. However other situations are full of covert activity. Driving for instance is one of them: the driver's attention and eye movements are mainly on the road, nonetheless he also pays attention to traffic lights and pedestrians that may cross the road. In the rest of this manuscript we consider the overt attention. More precisely, we aim to study towards which regions are the eyes are attracted.

3.1.4 The notion of "salient object"

While the first aim of saliency models consisted in predicting human fixations or scanpaths, a recent trend appeared with the work of Liu et al. [Liu 07, Liu 11] where they define the high level concept of "salient object" and propose a supervised approach where they learn to detect objects in an image.

Salient object detection or *Salient object segmentation* is commonly interpreted in computer vision as a process that includes two stages: 1) detecting the most salient object and 2) segmenting the accurate boundary of that object. The first stage does not necessarily need to be limited to one object. The majority of existing models have attempted to segment the most salient object, although their prediction maps can be used to find several objects in the scene.

The second stage falls in the realm of classic segmentation problems in computer vision but has certain differences (see Figure3.4).

Elazary and Itti [Elazary 08], analyzing LabelMe annotation data [Russell 08], demonstrate that human observers tend to annotate more salient objects first. They hence conclude that salient objects are interesting. Recently, Borji et al. [Borji 13b] conducted two experiments in which they asked 70 observers to explicitly choose the most outstanding (i.e., salient) object in a scene. In the first experiment, observers view scenes with only two objects. In the second experiment they ask observers to draw a polygon around the most salient object. These experiments revealed in particular that observers' judgments agree with saliency and eye movement maps.

Salient object detection has rapidly grown to become a full branch of saliency prediction and many studies have followed the original work by Liu et all [Liu 07, Liu 11] such as Achanta et al. [Achanta 08]. In more recent works [Li 14] Li et al. combine existing fixation-based saliency models with segmentation techniques in order to bridge the gap between fixation prediction and salient object segmentation.

A very thourough survey and benchmark on salient object detection have been conducted by Borji in respectively [Borji 12b] and [Borji 14].

3.2 Salient Features

This part of the manuscript describes the most known and used features regarding both top-down and bottom up models.

3.2.1 Bottom-up features

In this part we describe some features on which are based most bottom-up saliency maps models and present a brief history and description of the most-known bottom-up models.

3.2.1.1 Different features from the FIT

As previously mentionned, the FIT [Treisman 80] describe how stimuli can be decomposed into elementary attributes and which ones are important to attract bottom-up attention. Intensity, color, size, movement and orientation are the early features of the FIT [Treisman 80]. These features are traditionally used in bottom-up approaches as elementary cues to combine together in order to build a final master saliency map.

- Intensity or contrast are processed with algorithms inspired from LGN (Lateral geniculate nucleus) and V1 cortex center-surround neurons.
- Color is a useful cues, particularly when handling red-green and blue-yellow color opponencies. This is implemented by using color spaces such as HSV or LAB. Human perception of colors has been well studied by the team of C. Fernandez-Maloigne [Stoica 05, Larabi 09, Ivanovici 10].

- Orientation is usually computed by applying specific convolution processing or oriented Gabor filters.

- Motion is taken into account in most spatio-temporal saliency models (designed for video content). In fact, this is the relative motion which is salient from the HVS point of view. Usually, the visual scene relative motion is computed by applying a Global Motion Estimation (GME) methods such as in [Kraemer 04]. Daly in [Daly 98] noticed the saliency is not a linear function with the residual motion velocity. At some point, approximatly $30^{\circ}/s$, the eyes have difficulties to follow moving objects, and so saliency decrease to become null at $80^{\circ}/s$. An interesting study from G. Abdollahian et al. [Abdollahian 08] has explored the effects of camera motion on the visual attention. Their conclusion has stated that there is a high dependency between the camera direction and the gaze distribution. This means that the global motion has an impact on visual saliency as well.

- One other important cue is called the *central bias*. Many studies in bottom-up approaches for visual attention modeling have shown that the observers are attracted by the center of the stimulus ([Tatler 07, Dorr 10, Duan 11]). Indeed, the professional photographers and filmmakers use this phenomenon to place the most important objects and scene elements in the central position of a frame. In chapter 4 section 4.1 we describe the framework for bottom-up saliency computation introduced by Boujut et al. in [Boujut 12a] and make a contribution by studying the influence of various models of non-central bias based on recorded human data.

3.2.1.2 A brief history of bottom-up models

As previously mentioned, the feature integration theory has introduced how stimuli are decomposed in elementary attributes or cues which are important to attract attention. Here we present a few of the main models of visual attention inspired from this theory and the biology of the HVS.

The first model from Koch and Ullman It is in 1985 that Koch and Ullman propose the first plausible biologically inspired model of visual attention [Koch 85]. This model has been a milestone for many others since then such as the one from Itti and Koch [Itti 98]. According to Koch and Ullman, a set of visual features needs to be extracted from the original image in different features maps. These maps, following the processing in the visual cortex described in the FIT, decompose the stimuli in elementary cues such as colors, orientations,... These maps are then combined in a unique master saliency map encoding the whole saliency of the visual scene. A Winner Take All (WTA) network is then introduced to find out the most salient region in this master map - region called Focus of Attention (FoA). After the selection of the most salient region, the FoA is moved to a new region by discarding the previous FoA , and using once more the WTA network. This sequential process is done until finding a set of defined relevant zones.

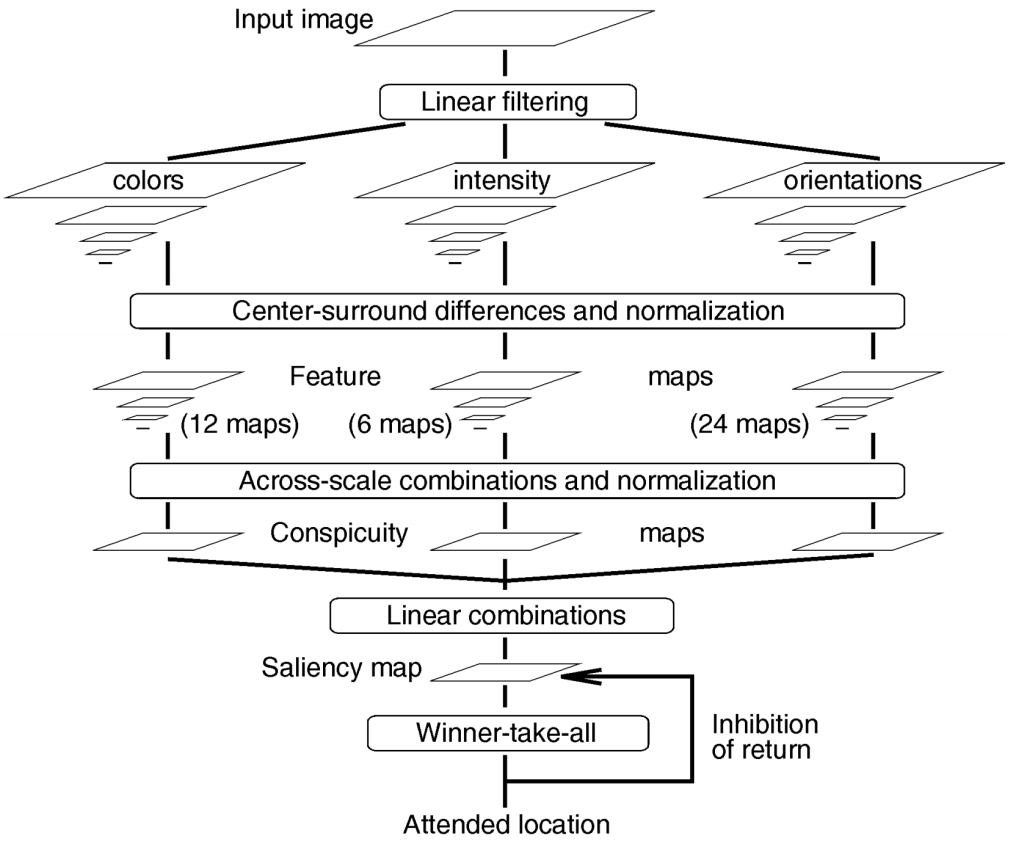


Figure 3.5 – General architecture of the Itti et al. model (image from [Itti 98])

The reference model from Itti, Koch and Niebur The visual attention model proposed by Itti et al. [Itti 98] in 1998, is beyond doubt the most famous model and the most cited. It has been used as a reference model in a lot of papers and is also used as such in this work. This model is based on the previously described architecture form Koch and Ullman [Koch 85] and is illustrated in Figure 3.5. Different attribute maps are extracted from the scene. They can be grouped in three main families described below.

- **An intensity map:**

$$I = (r + g + b)/3 \tag{3.1}$$

with (r, g, b) the red, green and blue channels of the image.

- **Four color maps:** a red R, green G , blue B and yellow Y map defined by:

$$R = r - (g + b)/2 \tag{3.2}$$

$$G = g - (r + b)/2 \tag{3.3}$$

$$B = b - (r + g)/2 \tag{3.4}$$

$$Y = \frac{r + g}{2} - \frac{| r - g |}{2} - b \tag{3.5}$$

- **Four orientation maps:** they are obtained from using Gabor filters (oriented pass-band filters) on the intensity map, aiming at representing the processing of the neurons in the in the primary visual cortex. The chosen orientations are 0°, 45°, 90° and 135°. In a second step, the attribute maps are transformed in *conspicuity maps* to give more weight on the regions which are different from their neighbors. Itti uses a multi-scale representation in order to be independent to the size of the salient region. Finally the maps from the same attributes family are combined to obtain a conspicuity maps by attributes (intensity, color, and orientation). This is done by a normalization operator $\mathcal{N}$ which normalizes each maps of the same family by the same values and then multiply each of them by

$$(M - \overline{m})^2 \tag{3.6}$$

with M the global maximum of the map and $\overline{m}$ the average of the other local maximums of the map. The goal of this normalization step is to give more importance to the maps with few peaks of high intensity value, corresponding to few salient zones, and to penalize the maps with a high number of peaks corresponding to uniform saliency values over the map. These normalized maps are then summed in a conspicuity map. Then the conspicuity maps are integrated to form a unique saliency map S. This integration process is illustrated in Figure 3.6.

$$S = \frac{1}{3}(\mathcal{N}(\overline{I}) + \mathcal{N}(\overline{C}) + \mathcal{N}(\overline{O})) \tag{3.7}$$

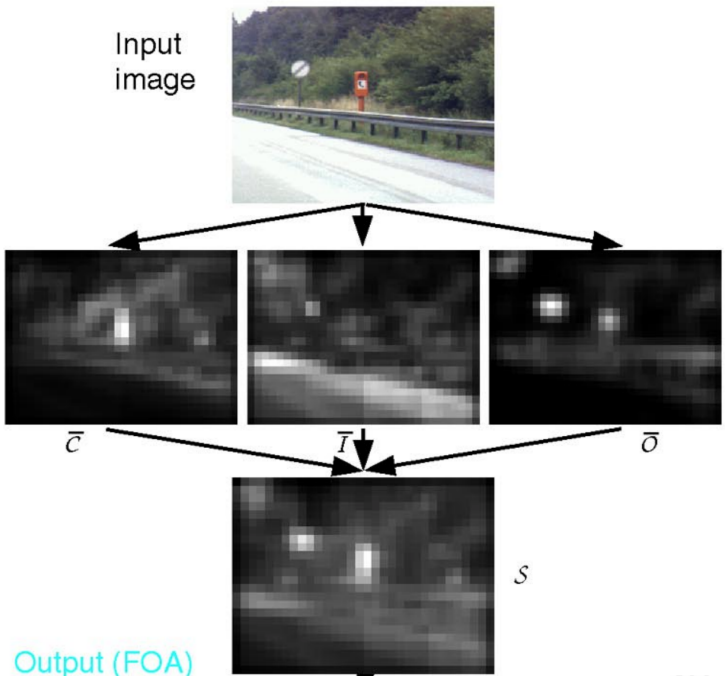


Figure 3.6 – Example of operation of the model with a natural image. Parallel feature extraction yields the three conspicuity maps for color contrasts $\overline{C}$, intensity contrasts $\overline{I}$, and orientation contrasts $\overline{O}$. These are combined the saliency map S (image from [Itti 98])

With respectively I, C and O the intensity, color and orientation conspicuity maps obtained before. The selection of salient zones then follow the model from Koch and Ullman with a WTA network that selects the most salient region, where is found the FoA, then sequentially discarding this region to find under the same WTA network the new FoAs.

And much more... Since then, much work has been done in this domain such as the well-known biologically inspired model from Le Meur also based on the FIT and on the architecture from Koch and Ullman. Another well-know bottom-up model is the Graph-Based Visual Saliency (GBVS) model from Harel et al. [Harel 07]. Later, in [Seo 09], Seo et al. proposed a model by extracting local features from an image which measure the likeliness of pixels to their surroundings. In [Brouard 09] , Brouard et al. introduced a method based on various color contrast descriptors computed in the HSV color space due to its closeness to human perception. The model proposed in chapter 4 is in fact a follow-up from the model of Brouard et al. Another well-known saliency model from Judd et al. [Judd 09] incorporated information from recorded eye-tracking data to the saliency computation. They propose a set of low, mid and high level image features to define salient locations and use a linear Support Vector Machine (SVM) [Cortes 95] to train the saliency model. The reader is referred to a recent benchmark of several saliency models for more details [Tilke 12].

Even though the majority of existing models are only applicable to static images, recent

studies have moved towards saliency in video. In [Marat 09], Marat et al. propose a method to fuse static an dynamic (relative motion) saliency cues with adaptive coefficients for each frames. Similarly the authors of [Zhong 13] merge classical bottom-up spatial saliency map with optical flow output based on dynamic consistency of motion. Rudoy et al. [Rudoy 13] describe a model that computes saliency by predicting the gaze location in a frame given the previous frame's fixation map. For the task of video surveillance, the authors of [Tong 11] introduced a background/foreground extraction step, to extract features (such as color, intensity, orientation but also faces detection) on the foreground regions. Their model then merges this static information with motion saliency maps. The applications of saliency maps to video content analysis are numerous: video quality assessment[Le Callet 01, Boujut 12c, LeMeur 10], gaze aware video compression [Komogortsev 09], video surveillance [Tong 11], activity recognition[González Díaz 13, Fathi 12], etc.

3.2.2 Top-down features and history

Although the literature concerning models of top-down attention is clearly less extensive than for bottom-up attention, the introduction of top-down factors (e.g., face, speech and music, camera motion) into the modeling of visual attention has provided impressive results in previous works [Ma 05, Cerf 07]. In addition, some attempts in the literature have been made to model both kinds of attention for scene understanding in a rather generic way. In [Huawei 14] the authors claim that the top-down factor can be well explained by the focus in image, as the producer of visual content always focuses his camera on the object of interest.

Recent works using machine learning approaches to learn top-down behaviours based on eye-fixation or annotated salient regions, have proved to be very useful for static images [Torralba 06, Gao 09, Kanan 09]. Furthermore, with the recent advances of Deep Learning Networks (DNN), some novel approaches have been designed in the field of object recognition, which build class-agnostic object detectors to generate candidate salient bounding-boxes which are then labeled by later class-specific object classifiers [Erhan 14, Shen 14]. However, it seems impossible for us to propose a universal method for prediction of the top-down visual attention component, as it is voluntary directed attention and therefore it is specific for the task of each visual search and very subjective. Nevertheless, the prior knowledge about the task the observer is supposed to perform, allows extracting semantic clues from the video content which would ease such a prediction.

The current state-of the art in computer vision allows detection of some categories of objects with a high confidence. A variety of face or skin detectors have been proposed since the last two decades [Jones 99, Viola 04]. Hence, when modeling a top-down attention in a specific visual search task, we can use such "easily recognisable" semantic elements that are relevant to the specific task of the observer and may help to identify the real areas/objects of interest.

In chapter 4 section 4.2 we propose a new top down probabilistic saliency model for egocentric video content. It aims to predict top-down visual attention maps focused on manipulated objects, that are then used for psycho-visual weighting of features in the problem of manipulated object recognition.

3.3 Human visual attention

In this section we present how human visual attention is measured and how to build visual attention maps from eye positions.

3.3.1 Recording fixations

The human visual attention is measured by observing the eye motion. Eye movements are depicted with a sequence of saccades, fixations, and smooth poursuits. Saccades are motions with a high amplitude that allow for the visual field exploration. On the contrary, fixations are micro-saccades with a low amplitude that place the object of interest over the fovea. Therefore, fine details are extracted during fixations. Smooth poursuits are triggered while tracking a moving object. Their role is to maintain the object over the fovea.

In 1967, A. L. Yarbus was the first to observe and describe with details the eye movements [Yarbus 67]. At that time, eye movement measurements were made with intrusive devices derived from contact lenses. Nowadays, eye position records are performed with non-intrusive devices called eye-trackers. These devices are made up with an infrared light and an infrared video camera. The infrared light illuminates the eye and the video camera records the eye. The infrared video camera and the two infrared lights are visible at the chinrest top on the Figure 3.7. The equipment on Figure 3.7 has an infrared mirror which only reflects the infrared light. With infrared lighting, the pupil appears dark on the video. A white spot is also visible over the pupil. It is the infrared light reflection on the eye. This spot is a landmark for the eye movement measurement. A digital processing is then required to track the white spot and the dark pupil. Eye movements are computed by comparing the centre coordinates from the spot and the pupil. In order to get accurate results, eye measures must be performed on several subjects under specific conditions as described for instance in the ITU-R BT.500-11 recommendation [ITU 02].

3.3.2 Building density maps from fixations

The subjective saliency maps in images and videos are built from eye position measurements in image/video plan. There are two reasons for which eye positions cannot be directly used to represent the visual attention. First, the eye positions are only spots on the frame and do not represent the field of view. Secondly, in order do deal with possible outliers, the saliency map is not built using only eye tracking data from one subject, but from many subjects. Therefore the subjective saliency map should provide an information about the density of eye positions. Today in various fields in vision research, such as computer vision, visual quality assessment in multimedia etc... the method proposed by D. S. Wooding [Wooding 02] has become the reference as it fulfills these two constraints. In the case of video sequences, the method is applied on each frame I of a video sequence. The process result is a subjective saliency map $S_{subj}(I)$ for each frame. With this method, the saliency map is computed in three steps. In the first step, for each eye measure m of frame I, a two dimensional Gaussian is applied at the center of the eye measure $(x_0, y_0)_m$. The two dimensional Gaus-

Figure 3.7 – Eye-tracker from Cambridge Research Ltd (picture from [Boujut 12b])

sian depicts the fovea projection on the screen. The fovea is the central retina part where the vision is the most accurate. In the *Sensibility to Light* [Hood 86], the authors state that the fovea covers an area from $1.5°$ to $2°$ in diameter at the retina center. D.S. Wooding propose to set the Gaussian spread σ to an angle of $2°$.

For the eye measure m of the frame I, a partial saliency map $S_{subj}(I, m)$ is computed such as:

$$S_{subj}(I, m) = Ae^{-\left(\frac{(x-x_{0_m})^2}{2\sigma_x^2}+\frac{(y-y_{0_m})^2}{2\sigma_y^2}\right)}$$
$$\text{with } \sigma_x = \sigma_y = \sigma \text{ and } A = 1 \tag{3.8}$$

Then, at the second step, all the partial saliency maps $S_{subj}(I, m)$ of frame $S_i(I)$ are added into $S_{subj}\prime(I)$:

$$S_{subj}\prime(I) = \sum_{m=0}^{N_I} S_{subj}(I, m) \tag{3.9}$$

where N_I is the number of eye measures recorded on all the subjects for the frame I. Finally, at the third step, the saliency map $S_{subj}\prime(I)$ is normalized by the highest value $argmax$ of $S_{subj}\prime(I)$. The normalized subjective saliency map is stored in $S_{subj}(I)$ The whole process is depicted in Figure 3.8.

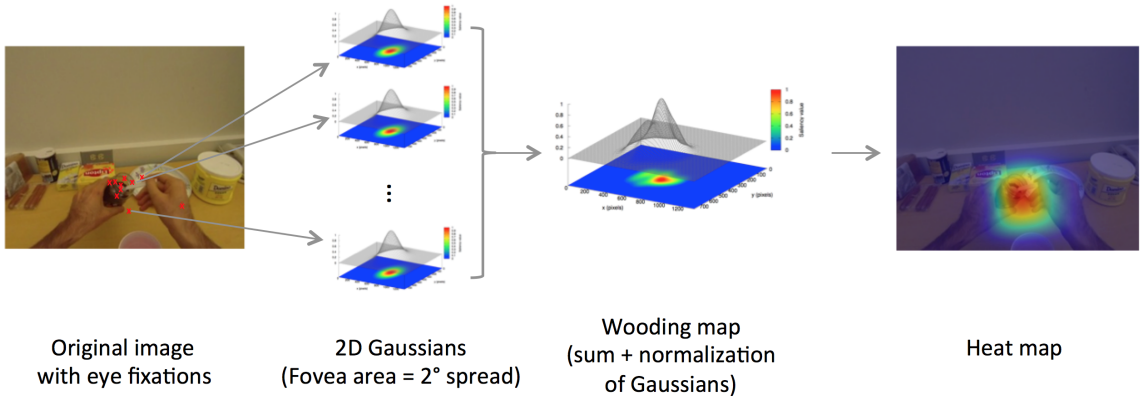


Figure 3.8 – Illustration of Wooding's method for creating visual attention maps from eye fixations

3.4 Evaluation metrics

In this part of the manuscript we introduce a few methods to assess the validity of automatic saliency maps. The first set of methods is based on the comparison with human recordings of eye positions while the second aims at measuring how well saliencyi maps perform in the framework they were developed for (i.e. object segmentation, object recognition, ...).

3.4.1 Comparing with fixation points

In the literature, many metrics have been used to compare automatically generated saliency maps with recorded eye-fixation maps from humans observers. Among these metrics we can list the Kullback-Leiber divergence (KLB) [Tatler 05, Le Meur 07, Peters 08], the Area Under Receiver Operating Curves (AUC) [Tatler 05, Judd 09, Le Meur 07], the Pearson Correlation Coefficient (PCC) [Le Meur 07, Ouerhani 03], the Normalized Scanpath Saliency (NSS) [Parkhurst 02, Peters 05], the parsing diagrams, the percentage of fixations in salient zones, etc ... For the experimental sections of this manuscript where such metrics have been needed, we retained three in total: the PCC, the Area Under Curve (AUC) and the NSS. We provide a definition of these three metrics below.

3.4.1.1 PCC

The PCC, in this context, gives information about the existence of a linear relationship between the automatic saliency model S_a and the ground truth human fixation maps S_h. It is defined by

$$PCC(S_a, S_h) = \frac{cov(S_a, S_h)}{\sigma_{S_a}\sigma_{S_h}} \tag{3.10}$$

with σ_{S_a} (respectively σ_{S_h}) the standard deviation of the values in the map S_a (respectively S_h) and $cov(S_a, S_h)$ the covariance between S_a and S_h. This coefficient is bounded

between $[-11]$. A value equal to zero proves the absence of linear relation between these two maps: there is no correspondence between automated saliency and eye fixations. The closer the value is from 1, the more high saliency value in S_a correspond to the zones the most looked at (high values in S_h). Oppositely the closer the value is to -1, the more high saliency value in S_a correspond to the zones the least looked at (low values in S_h).

3.4.1.2 AUC

AUC is a popular metric in the research community as well. It is suitable only for pixel-based saliency map. This method is proposed in Signal Detection Theory and Psychophysics published in 1966 by D. Green and J. Swets [Green 66]. It is applied to assess binary classifiers. In the case of saliency maps, a threshold should be used to distinguish salient from non-salient regions. This metric is easy to implement, nonetheless, it requires high computational resources if several thresholds are tested. The two continuous maps: the ground truth S_{subj} and predicted from image signal S_{obj} maps are thresholded. The S_{subj} is thresholded with a constant threshold in order to keep a given percentage of top salient pixels. In [LeMeur 12] the authors propose to keep the top 2%, 5%, 10% or 20%. The corresponding threshold is called TG_x (G for ground truth and x indicating the percentage of image considered as being fixated). The threshold for S_{obj} is systematically moved between the minimum and the maximum values of the map. For each pair of thresholds the number of true positives (TP), the false positives (FP), the false negatives (FN), and the true negatives (TN) are computed. The true positive number is the number of fixated pixels in the ground truth that are also labeled as fixated in the prediction. Then the ROC curve that plots the FP rate $FPR = \frac{FP}{TP+FN}$ as a function of TP rate $TPR = \frac{TP}{TP+FN}$ is used to display the classification result. The AUC provides a measure indicating the overall performance of the classification.

3.4.1.3 NSS

NSS is a Z-score that compares two scanpaths. This measure has been first introduced by D. Parkhurst et al. [Parkhurst 02] and R. Peters et al. [Peters 05]. Later, the NSS has been modified by A. Bur et al. [Bur 07] to compare two pixel-based saliency maps. This method is widely used in the research community since it is suitable for scanpath and pixel-based saliency maps. Only few computational resources are required which makes this metric well-suited for the evaluation of saliency on videos. We use the NSS metric that has been adapted in [Bur 07] to pixel-based saliency. NSS is a Z-Score that expresses the divergence of human visual attention maps from gaze measures, from the objective saliency maps extracted from image signal. The NSS computation for a frame I is depicted by equation (3.11).

$$NSS = \frac{\overline{S_{subj} \times S_{obj}^N} - \overline{S_{obj}}}{\sigma(S_{obj})} \tag{3.11}$$

Here, S_{obj}^N denotes the objective saliency map S_{obj} normalized to have a zero mean and a unit standard deviation, $\overline{X}$ means an average. When $S_{subj} \times S_{obj}^N$ is higher than the average objective saliency, the NSS is positive; it means that the gaze locations are inside the saliency

depicted by the objective saliency map. In other words, the higher the NSS score, the higher the similarity between compared saliency maps.

3.4.2 Metrics by recognition performances

Comparing saliency maps with human visual attention maps thanks to the metrics presented before in 3.4.1 is not the only way to evaluate the validity of saliency maps. It is also possible to evaluate directly how well they perform in the framework for which they were designed. In our case, since our goal is to design saliency maps aiming at recognizing active objects in egocentric videos, it is possible to base the quality metric for the maps on active object recognition performances. This will be the case in chapters 4 and 5.

3.5 Conclusion

In this chapter we first provided the important definitions and characteristics about visual attention modeling. Then we introduced a non-exhaustive list of salient features, in particular, we talked about bottom-up features and top-down ones. In order to build a ground truth, we then presented a method to build human visual attention maps from gaze recordings. Finally we introduced ways to evaluate how visual attention models are compared with human visual attention and how they can be evaluated in a task-specific framework such as active object recognition. In the next chapter we will present our contribution in saliency modeling both for the bottom-up and top-down domains.

Chapter 4

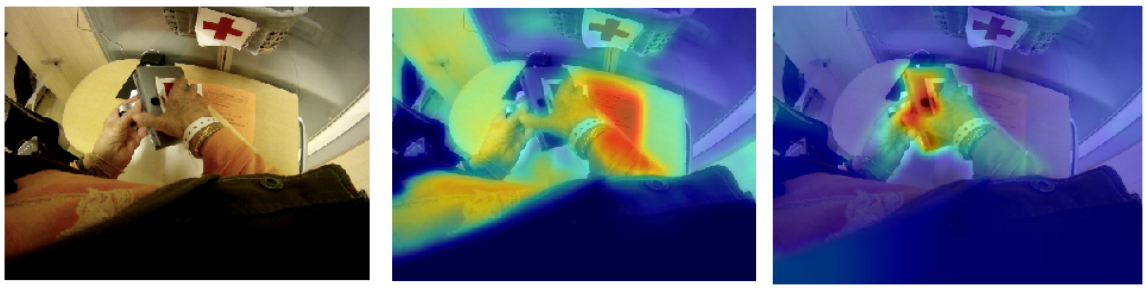
Saliency maps for object recognition in egocentric video

In this chapter we investigate and propose a contribution in both Bottom-up (gaze driven by stimulus) and Top-down (gaze driven by semantics) areas that aim at enhancing the particular task of active object recognition in egocentric video content. Our first contribution on Bottom-up models takes its roots from the fact that observers are attracted by a central stimulus (the center of an image). This biological phenomenon is known as *central bias*. In egocentric videos however this hypothesis does not always hold. We study saliency models with non-central bias geometrical cues. The proposed visual saliency models are trained based on eye fixations of observers and incorporated into spatio-temporal saliency models.

Regarding the top-down domain, we present a probabilistic visual attention model for manipulated object recognition in egocentric video content. Although arms often occlude objects and are usually seen as a burden for many vision systems, they become an asset in our approach, as we extract both global and local features describing their geometric layout and pose, as well as the objects being manipulated. We integrate this information in a probabilistic generative model, provide update equations that automatically compute the model parameters optimizing the likelihood of the data, and design a method to generate maps of visual attention that are later used in an object-recognition framework.

4.1 Contribution to Bottom-up saliency prediction

Many studies in bottom-up approaches for visual attention modeling have shown that the observers are attracted by the center of the stimulus ([Tatler 07, Dorr 10, Duan 11]) as introduced in section 3.2.1. This phenomenon is known as *center bias*. Indeed, the professional photographers and filmmakers use this phenomenon to place the most important objects



(a) (b) (c)

Figure 4.1 – Illustration of visual saliency cues: from left to right, original frame, spatial, temporal

and scene elements in the central position of a frame. In egocentric videos, which have not undergone a video production process, this hypothesis does not hold. The locus of important visual information depends on the camera position on a body. In [Boujut 12a] the authors propose a geometrical model which is trained on the basis of a psychovisual experiment. Subjects observed egocentric video content recorded with a shoulder-mounted cameras and their gaze fixations were recorded. In this part of the manuscript we further study saliency models with non central bias geometrical cues developing the idea of [Boujut 12a]. The proposed visual saliency models are trained based on eye fixations of observers and incorporated into spatio-temporal saliency models. We assess these new models in terms of pattern recognition performances and by comparing to state of the art models widely used for still images [Itti 98, Harel 07] and videos [Seo 09] by the mean of dedicated metrics 3.4.1.1. The results are promising:they highlight the necessity of a non-centered geometric saliency cue.

The rest of the part of this section is organized as follows: section 4.1.1 presents spatio-temporal visual saliency models with geometric cues. Experimental set-up and details as well as the dataset are described in section 4.1.2. The experimental results for the different saliency models are presented and discussed in section 4.1.3 and section 4.1.4. Conclusions and perspectives are given in section 4.1.5.

4.1.1 Visual saliency models for object recognition

The overall model of visual attention predictor proposed in the present work is a follow-up of the model [Boujut 12a]. The latter itself was based on the previous research of O. Brouard et al. [Brouard 09]. It explores residual motion in the scene, color and spatial contrasts, and the so-called central-bias hypothesis which means, that the human gaze is attracted by the center of the image. In [Boujut 12a] they extended the state of the art specifically modelling geometrical saliency and proposed multiple fusion schemes for the temporal (fig. 4.1c), spatial (fig. 4.1b) and geometrical cues. In the following we briefly present each component and focus on our contribution.

Spatial saliency S_s: proposed in [Brouard 09], is based on various color contrast descrip-

tors that are computed in the HSV color space, due to its closeness to human perception of color. In particular, 7 local contrasts are computed, namely:

1. *Contrast of Saturation*: A contrast occurs when low and highly saturated color regions are close.
2. *Contrast of Intensity*: A contrast is visible when dark and bright colors co-exist.
3. *Contrast of Hue*: A hue angle difference on the color wheel may generate a contrast.
4. *Contrast of Opponents*: Colors located at the hue wheel opposite sides create very high contrast.
5. *Contrast of Warm and Cold Colors*: Warm colors – red, orange and yellow – and cold ones such as blue neighboring in an image create visually appealing contrasts.
6. *Dominance of Warm Colors*: Warm colors are always visually attractive even if no contrast are present in the surrounding.
7. *Dominance of Brightness and Saturation*: Highly bright and saturated regions have more chances of attracting the attention, regardless of the hue value.

The spatial saliency value $S_s(i)$ for each pixel i in a frame is computed by averaging the outputs associated to the 7 color features.

Temporal saliency S_t: In the present work we base our temporal cue on residual motion employing the model from [Liu 09]. This saliency models the attraction of attention to motion singularities in a scene. Human visual attention is not grabbed by the motion itself, but by the residual motion for each pixel, e.g. the difference between the estimated motion for each pixel and the predicted camera motion based on a global parametrization.

Omitting many details, the process of computing a temporal saliency map is as follows: first, for each frame in the video, a dense motion map $\mathbf{v}(i)$ that contains the motion vectors at each pixel i in the image is computed using the optical flow method described in [Farnebäck 00].

Then, a 3x3 affine matrix A that models the global motion (GME) associated to the camera movements is computed. For that end, the well known robust estimation method RANSAC [Fischler 81] has been used in order to successfully handle the presence of outliers (e.g. areas of the image associated to objects that move differently than the camera). Furthermore, since the central area of each frame constitutes the most likely region where moving objects appear, this region is not considered for the affine matrix estimation, thus reducing the proportion of outliers.

Next, the residual motion $\mathbf{r}(i)$ is computed by compensating the camera motion:

$$\mathbf{r}(i) = \mathbf{v}(i) - A\mathbf{x}_i \tag{4.1}$$

where $\mathbf{x}_i$ stands for the spatial coordinates of each pixel i, $\mathbf{x}_i = (x_i, y_i, 1)^T$ and A is the 3x3 affine matrix modeling global motion (GME).

Finally, the values of the temporal saliency map $S_t(i)$ are computed by filtering the amount of residual motion in the frame. The authors of [Brouard 09] reported that the human eye cannot follow objects with a velocity higher than $80°/s$ [Daly 98]. According to this psycho-visual constraints, a post-processing filter was proposed in [Brouard 09] that decreased the saliency when motion was too strong. Applying this filtering stage to our first-person camera videos was however too restrictive due to the strong camera motion so that we have preferred to consider a simpler filtering stage that normalizes and computes the saliency map as follows:

$$S_t(i) = \min\left(\frac{||\mathbf{r}(i)||_2}{K}, 1\right) \tag{4.2}$$

where K has been heuristically computed depending on image dimensions (H,W), as $K = \max(H, W)/10$.

Geometrical saliency S_g: We pay a particular attention to the geometrical cue. In [Brouard 09], the geometric saliency map $S_g(i) = \mathcal{N}((x_0, y_0), (\sigma_x, \sigma_y))$ is computed as 2D Gaussian located at the screen center with a spread $\sigma_x = \sigma_y = 5°$. This model expresses the central bias hypothesis [Buswell 35]. However, this affirmation only holds for videos coming from professional and hand-carried cameras such as in movie datasets [Vig 12]. Indeed based on previous experiments of [Boujut 12a] who recorded eye positions of 21 subjects looking at egocentric content with instrumental activities in view field and shoulder mounted camera(see figure 4.2), one can see that a central-bias hypothesis does not hold anymore here. The center of fixations is notably higher than the center of the frame and even the shape itself is more elliptical than circular. Indeed, with such a camera setting, the center of the frame is not necessary the place where people look at: instead, they focus on activities. Depending on the body-worn camera position, the area of activities can be shifted. Based on the observations of [Boujut 12a], we explore in this part of the manuscript three different geometrical saliency cues:

- Geometrical centered circular: this is a circular centered uniform Gaussian with a spread of σ_x = σ_y = 5 visual degrees. This is a baseline expressing central-bias hypothesis. For the rest of this study, the spatio-temporal visual saliency model using this geometrical cue will be referred as STC (see figure 4.3a).
- Geometrical not-centered circular: this cue is similar to STC but the center has been shifted in order to match the one computed from the mean horizontal and vertical positions of the eye fixations of observers (see figure 4.2). For the rest of this study, the spatio-temporal visual saliency model using this type of geometrical cue will be referred as STNC (see figure 4.3b).
- Geometrical not-centered flared: this last type of geometrical saliency has the same center as the geometrical not-centered circular but σ_x and σ_y are unequal and computed from Gaussian fitting on eye fixations (see figure 4.2). For the rest of this study, the spatio-temporal visual saliency model using this geometrical cue will be referred as STNCF (see figure 4.3c).

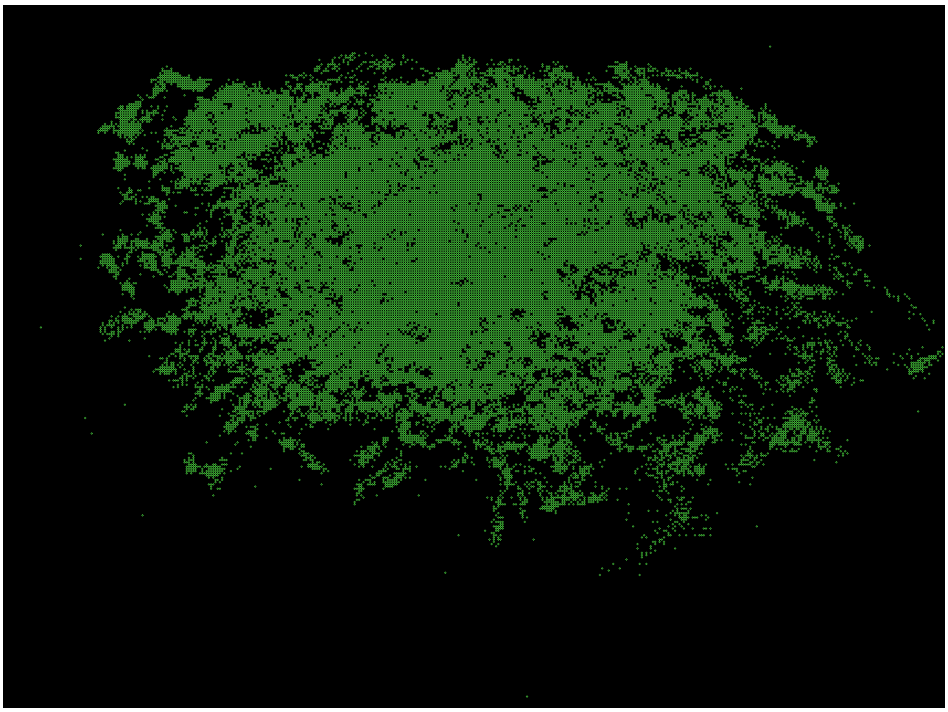


Figure 4.2 – Superposition of fixation points from 15 observers and 3 videos recorded by the shoulder-mounted GoPro wearable camera

Fusion of Saliency cues: Compared to [Boujut 12a], we also use a new fusion scheme for the temporal, spatial and geometrical cues. Denoting respectively these three cues by s, t, g , we compute the fusionned saliency map by a linear combination, see eq.4.3.

$$S(i) = w_s S_s + w_t S_t + w_g S_g \tag{4.3}$$

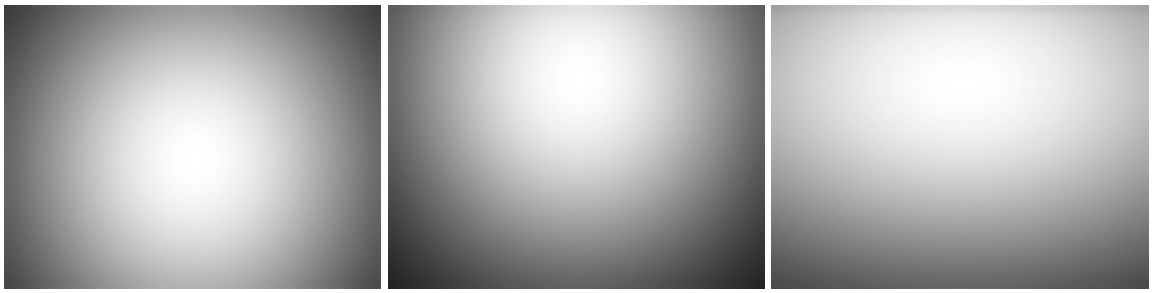
Here w_j are the weights associated with the jth saliency map ($j = s, t, g$). The coefficients are learned by regression on a sub-dataset consisting of randomly selected frames from all the training videos (to avoid loss of generality) with bounding boxes of objects as ground-truth. On the contrary to [Boujut 12a], where ad-hoc fusion schemes are used, such a method adapts saliency maps to a video corpus in a machine learning framework. The values of $S(i)$ are then normalized within the interval $[0, 1]$ for each frame.

4.1.2 Experimentation protocol

In this section we introduce egocentric video dataset and then introduce other state-of the art saliency models for comparison.

4.1.2.1 Dataset

The dataset we used in this context was recorded in a hospital environment for the sake of the Dem@care project. It consists of egocentric videos of patients at early stage of Alzheimer



(a) Geometrical centered circular (b) Geometrical not centered circular (c) Geometrical not centered flared

Figure 4.3 – Illustration of the three geometrical saliency models integrated to the final spatio-temporal-geometric one

disease performing instrumental activities of daily living in a hospital environment. The videos were recorded by a shoulder-mounted GoPro camera with the resolution of 1280×960 pixels at frame-rate of 30 fps. The recorded dataset was composed of 44 videos of approximate duration of 9 hours and 30 minutes overall. In this corpus 22236 frames were manually annotated and a taxonomy of 17 active objects' categories was defined. The objects were annotated in all the videos delimiting them with bounding boxes. The dataset was finally divided into a training and test sets of videos. The split was optimized to even the number of objects for each category in both sets. The number of object instances varied across the categories from 102 ("tea box") to 2032 ("tablet").

As in the present work the camera position on the body was the same as in [Boujut 12a], a cross-corpus training of geometrical cues was possible. We thus used the coordinates of fixations of human observers (see section 4.1.1) from [Boujut 12a] as training data for estimating the parameters of our geometrical models STNC and STNCF.

4.1.2.2 Setting up the final model

Here we explain the computation of the coefficients for the linear combination of spatial, temporal and geometrical cues, see section 4.1.1. Coefficients were computed by regression over a sub part of the training dataset. More exactly, random images from the training dataset were selected in all videos for the sake of generality. We computed the coefficients from equation 4.3 so that the linear combination matches the locations of manually annotated bounding boxes. The computed coefficients are given in table 4.1 for all three models.

It can be seen that, the weights for the STNC model are very much unequal (see line 2). Indeed, the weight w_g is much higher for STNC and STNCF compared to the weights of the temporal and spatial cues. The rationale behind this is that the not-centered geometrical map (see figure 4.3b) explains well where bounding boxes are mostly situated. To illustrate this, we created a normalized map by superimposing all the bounding boxes over a small random part of the training dataset. The resulting "Bounding Box (BB) map" is displayed in figure 4.4.

Table 4.1 – Coefficients of the linear combination of spatial, temporal and geometrical cues for the three different types of geometrical maps

	w_s	w_t	w_g
STC	0.36	0.29	0.35
STNC	0.05	0.1	0.85
STNCF	0.29	0.26	0.46

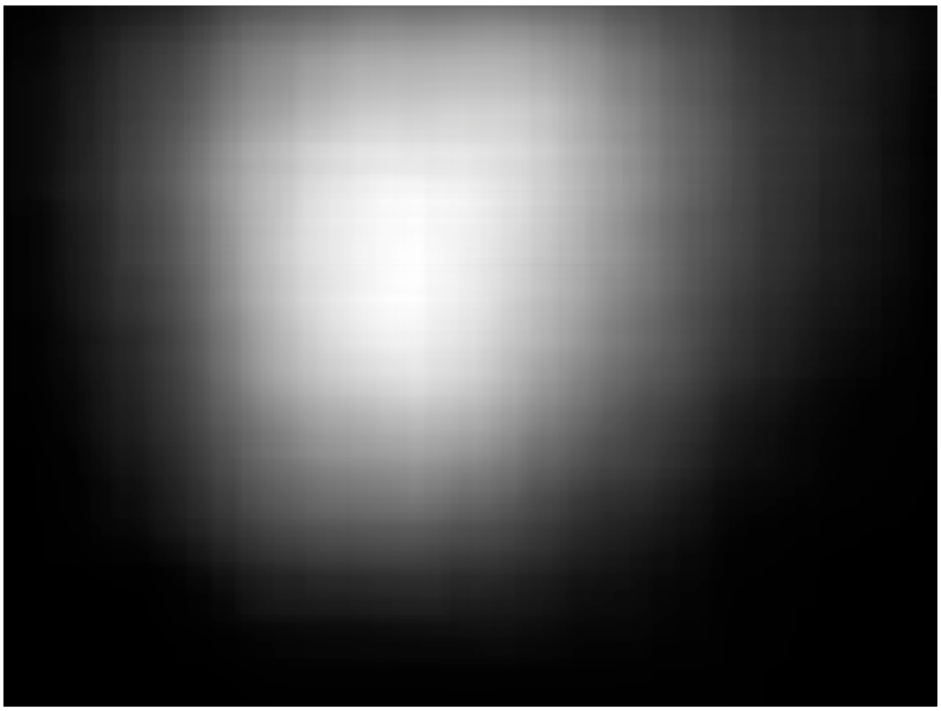


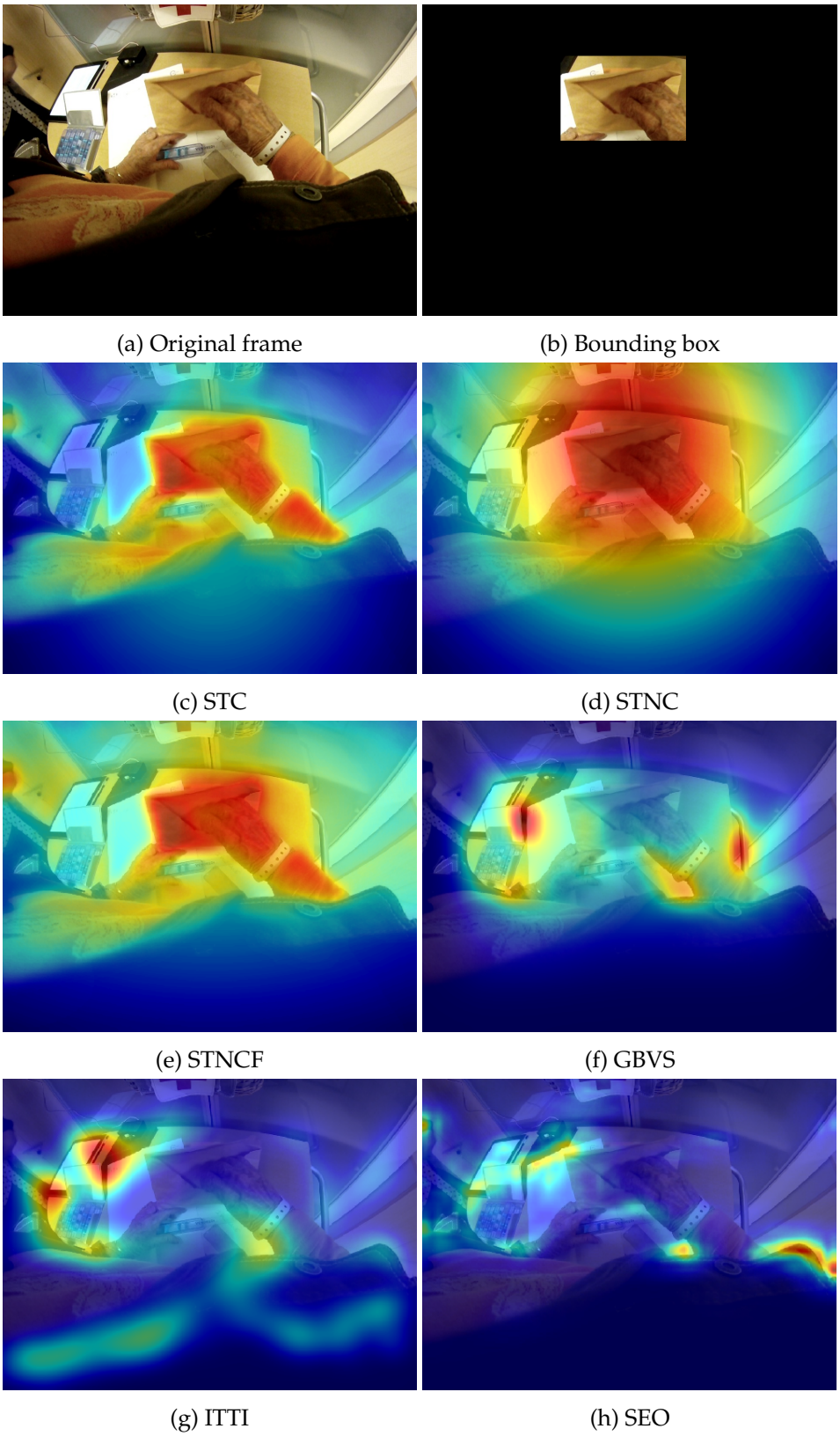
Figure 4.4 – Superposition of manually annotated bounding boxes

The superposition of objects' locations (as illustrated in figure 4.4) tends to show a high similarity with the not-centered Gaussian cue (figure 4.3b). This supports our assumption: the central hypothesis for a Gaussian does not fit this kind of egocentric video content.

4.1.2.3 Selected Saliency models for comparison

For the experimentations, we first chose the three models introduced in section 4.1.1, i.e. STC, STNC, and STNCF. Our models are compared with 3 widely used saliency models:

- Reference model of Itti [Itti 98]. We will refer to this model as "ITTI" later on.
- The graph-based visual saliency model developed by Harel [Harel 07]. It will now be referred to by the acronym "GBVS".



(a) Original frame (b) Bounding box

(c) STC (d) STNC

(e) STNCF (f) GBVS

(g) ITTI (h) SEO

Figure 4.5 – Original frame, manually annotated bounding box, and the different saliency models selected for this study as heat maps

- Since our models introduce the notion of temporality, we also chose a model based on spatio-temporal ressemblance, developed by H. J. Seo [Seo 09]. It model was chosen since it has been part of a recent study about gaze prediction to identify objects [Borji 13a]. This model will be called "SEO" for the rest of this study.

The ITTI and GBVS models have been selected in this comparison since they are frequently used in literature for computing saliency maps.

Figure 4.5 illustrates computed saliency maps for a randomly selected video frame. We also display the manually annotated bounding box corresponding to the active object ("envelope").

For the sake of comparison, we also defined our baseline model as the one without any saliency maps. This is a conventional BoVW approach with a dense sampling of features on the whole frame. The latter will be referred as "Simple BoVW" for the rest of this study. For the computation of BoVW, we used a dictionnary size of 4000 visual words.

4.1.2.4 A quality metric for saliency models based on object recognition performances

Since our goal is to find saliency maps for an active object recognition task, we base the quality metric for the maps on object recognition performances as presented in 3.4.2.

In our work, we used the well-known BoVW paradigm 2.3.1.1 combined with a saliency pooling step as will be introduced in 5.2.1. Simply put, features over salient areas will get more weight in the image signature than features over non-salient areas. In the present work, we use the 64-dimensional SURF features [Bay 08]. Once each image is represented by its weighted histogram of visual words, we use a non-linear classifier to detect the presence of a category in the image. In particular, we have employed a SVM classifier [Cortes 95] with a χ^2 kernel, which has shown good performances in visual recognition tasks working with normalized histograms. The target quality metric is the Average Precision (AP) for all objects categories. The mAP of all the categories is also been computed.

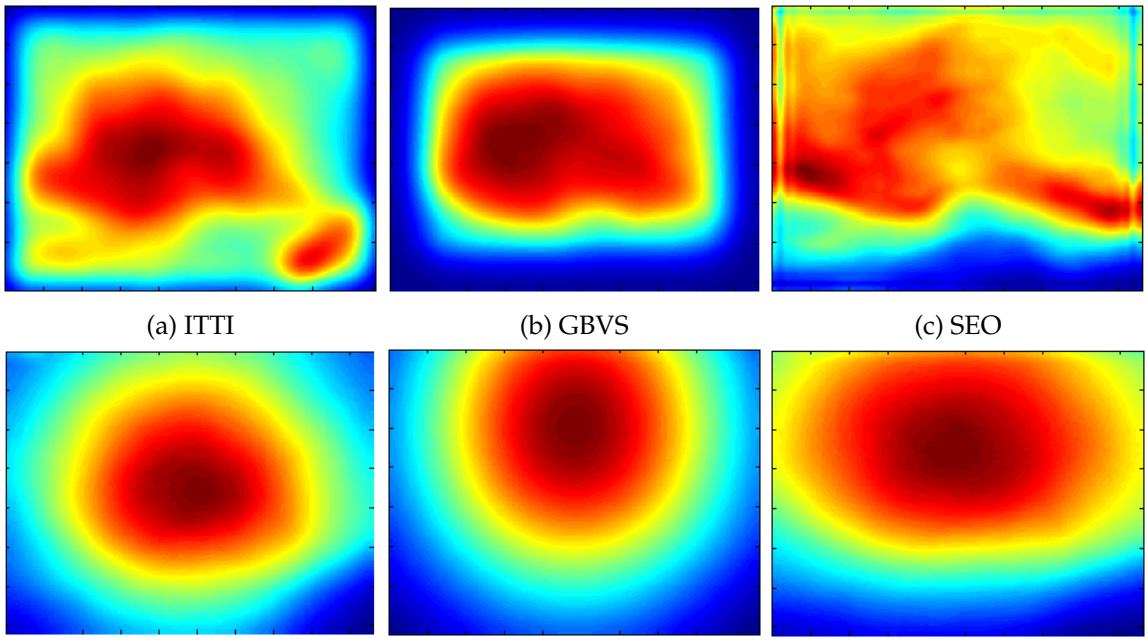
4.1.3 Psycho-visual evaluation of proposed saliency models

Before evaluating our geometrical saliency cues with the object recognition metric presented in section 4.1.2.3, we want to assess their capacity to predict human visual attention.

Automatically predicted saliency maps can be compared to human gaze fixations with help of dedicated metrics. From [Riche 13] and anterior work [LeMeur 12] we retained the PCC.

As can be seen in section 3.4.1.1 a value close to 1 means the good correspondance of saliency maps.

First of all, we compute the PCC between the superposition of BB map presented in fig.4.4 and the superposition of all corresponding saliency maps (for ITTI, GBVS, SEO, STC, STNC, and STNCF see fig.4.6) to assess how the saliency maps main highlighted locations concord with the average objects of interest locations. Table 4.2 display the results of this



(a) ITTI (b) GBVS (c) SEO

(d) STC (e) STNC (f) STNCF

Figure 4.6 – Superposition of saliency maps in all annotated frames (similarly to fig.4.4) for the six different types of automatic saliency models chosen in this study

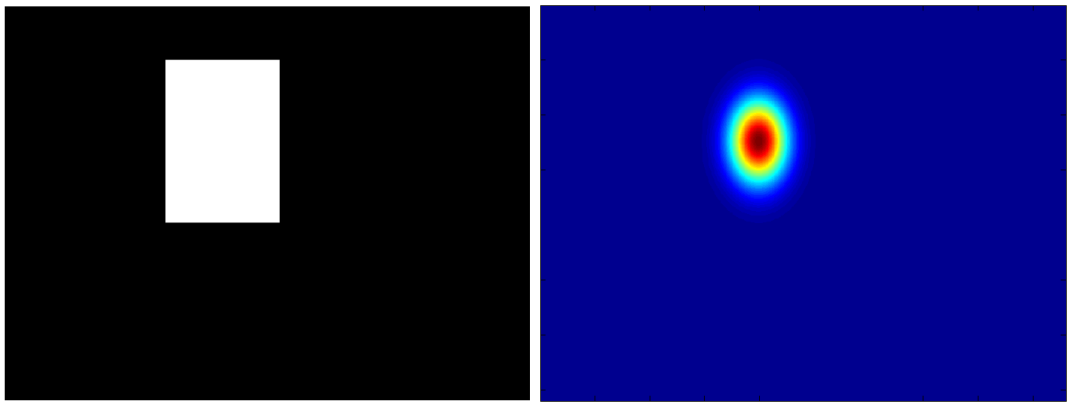
Table 4.2 – Mean PCC between the superposition for all annotated frames of BB maps and the different considered saliency maps

	ITTI	GBVS	SEO
PCC	0.4743 ± 0.3032	0.6633 ± 0.1905	0.5207 ± 0.2096
	STC	STNC	STNCF
PCC	0.7850 ± 0.1253	0.9093 ± 0.0808	0.7770 ± 0.1303

rough PCC evaluation. We can conclude that the mean PCC is higher for all our spatio-temporal-geometric models and more precisely that STNC gives the highest correlation coefficient.

Since the dataset contains 22236 annotated frames we could not perform a guided psycho-visual experiment to measure eye fixations. We considered instead the ground truth bounding boxes to be reference visual attractors. Based on this assumption, fitted gaussians centered on geometrical centers of BB are then considered as simulated human visual attention maps (see fig. 4.7) with which were conducted the psycho-visual evaluation.

We measured the similarity of recorded generated eye fixations and automatically generated saliency maps from our spatio-temporal-geometric model using the three different geometrical cues (STC, STNC, STNCF) and the ones presented in section 4.1.2.3. In total 22236 frames were compared for each saliency model and the final mean scores with stan-



(a) Bounding Box (b) Generated human visual attention map

Figure 4.7 – Illustration of the generated visual attention map from bounding boxes

Table 4.3 – PCC mean scores (with standard deviations) between generated human fixation points and different saliency map models.

	ITTI	GBVS	SEO
PCC	0.1417 ± 0.1254	0.1752 ± 0.1195	0.1151 ± 0.1093
	STC	STNC	STNCF
PCC	0.1987 ± 0.1778	0.2239 ± 0.1308	0.1854 ± 0.1404

dard deviations are presented in table 4.3.

First we can see that all the scores obtained with our spatio-temporal-geometric models are higher than ITTI, GBVS and SEO saliency models. Since the standard deviation are high, we computed the p-values to back up the hypothesis that the spatio-temporal-geometric models are significantly higher than with the other saliency models. At the 5% significance level, the data do provide sufficient evidence to conclude that the mean PCC score using our spatio-temporal-geometric models are greater than the mean obtained using other saliency models (p-values are of the order of 10^{-13} in all cases). Finally the same conclusion can be drawn that for table 4.2 that is to say among the three geometrical cues in our spatio-temporal-geometric saliency model, the not centered geometrical cue (STNC) gives a higher average response value to our generated human eye positions.

4.1.4 Object recognition results

We assessed our saliency models against the benchmark specified in section 4.1.2.3. The object recognition approach and the quality metric are those presented in section 4.1.2.4.

Object recognition results are illustrated in figure 4.8 in terms of AP (Average Precision). They show the detailed, per-category performances for the different selected models of saliency. The last column is the mAP computed on all categories. Mean Average Precision

Table 4.4 – mAP for the referenced and proposed saliency models computed on the first nine categories of figure 4.8.

	BoVW	GBVS	ITTI	SEO
mAP	0.77 ± 0.20	0.74 ± 0.21	0.73 ± 0.24	0.75 ± 0.21

	STC	STNC	STNCF
mAP	0.78 ± 0.21	0.77 ± 0.21	0.79 ± 0.20

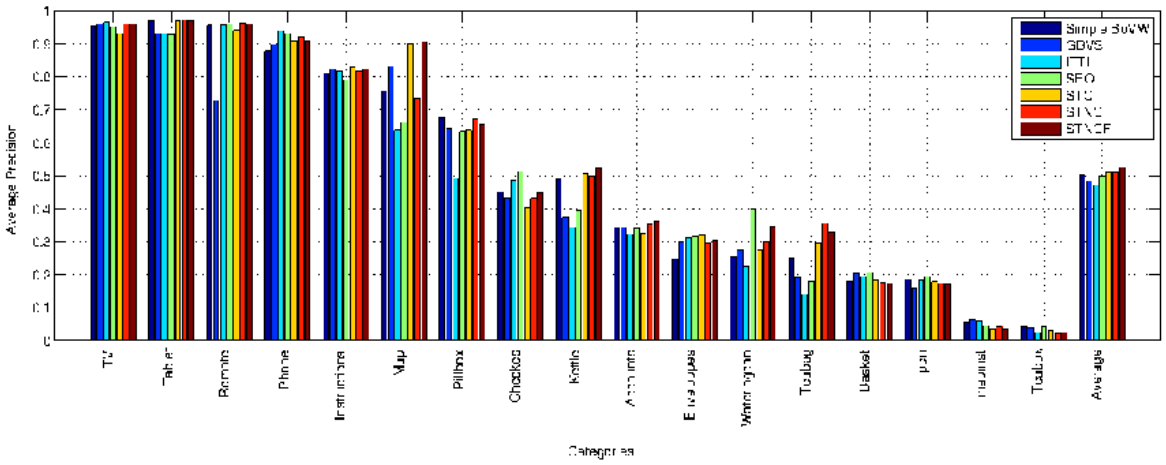
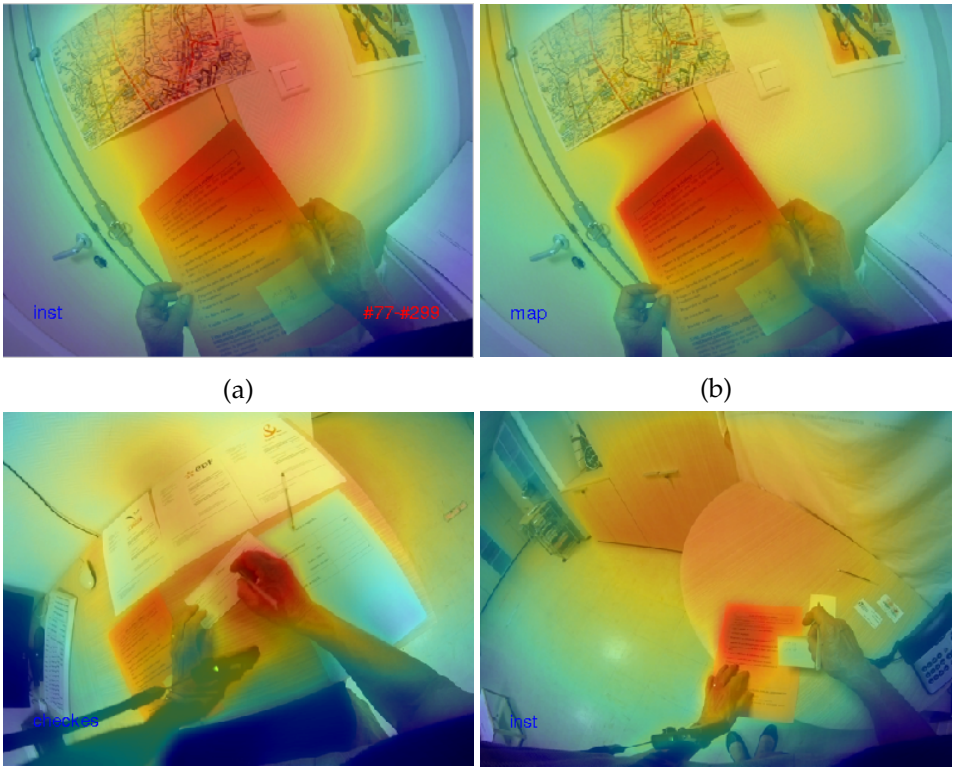


Figure 4.8 – Object recognition results for the different selected saliency models



(a) (b)

(c) (d)

Figure 4.9 – Illustration of the different kinds of errors observed: a) STNC, b)STNCF, c)misrecognition of "checks", d)missrecognition of "cards"

values for all categories of objects vary from 0.48 for Itti model to 0.51 for STNC (see the last column in figure 4.8 for illustration) with rather large variance of 0.32. This can be explained by the presence of categories where mAP is generally low as the objects are small in image plane. The typical kinds of errors found are displayed in figure 4.9 such as: 4.9c. spatial proximity and similarity of features,4.9d. low number of annotated objects and small size.

For the specific case of the object "map", the performances of the STNC model (AP of 0.73) are lower than those of the STNCF and STC models with both APs of 0.9 (see figure 4.8). The saliency predicted by the the STNC model is more focused on the map located in the centered higher part of the frame whereas the STNCF focuses on the real manipulated object: "instructions" (see figures 4.9a, 4.9b), leading to false positives. We present in table 4.4 statistics on the categories for which at least one AP value is above 0.5 (see also recognition results illustrated in fig.4.10 for clarity). Note, that from object recognition perspective the STNCF model performs the best, despite it is less correlated with the BB map than STNC. This can be explained by the better coverage of significant features in image plane.

Based on these results, we can draw the following conclusions:

- The models we propose perform better (or comparably for STNC on the "best categories") than baseline BoVW paradigm. It is, however, not the case with the other saliency models. ITTI, GBVS, and SEO saliency maps are too sparse (see figures 4.5g,

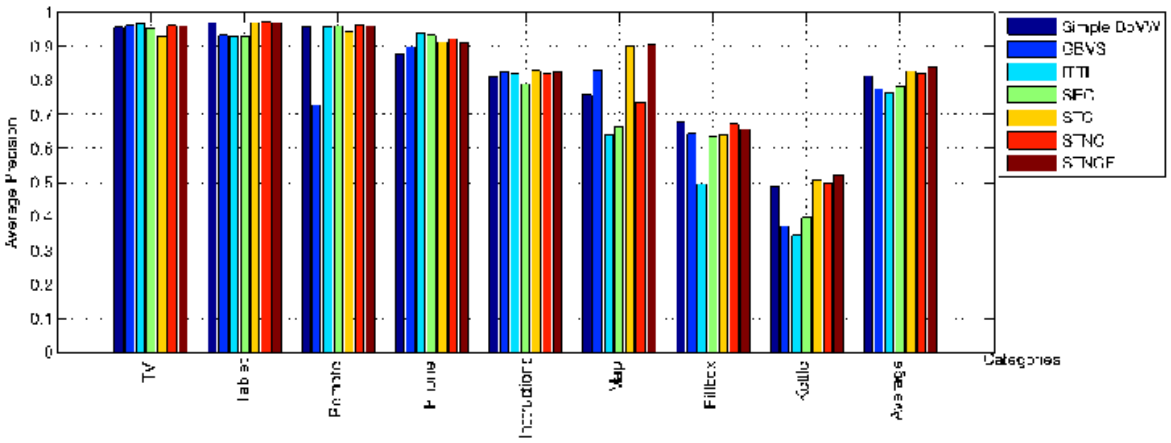


Figure 4.10 – Object recognition results for the categories with simple BoVW performance is above 0.5

4.5f, 4.5h) and consequently, do not cover the objects of interest enough. Also, none of these methods include a geometric cue.

- Among our three models, the STNCF yields better performances than STNC. This observation confirms our initial hypothesis. The geometrical cue is indeed important but it needs to be adapted to the video content.

Learning the coefficients for the linear combination of spatial, temporal and geometrical cues in our saliency model (see section 4.1.2.2) is necessary. Indeed, we computed saliency maps with equal coefficients (0.33), using mean pooling as tested in [Boujut 12a] for the STNCF and the mAP resulted in a few percent lower values than those reported in figure 4.8.

4.1.5 Discussion and perspectives

In this first contribution we have presented a method for building bottom-up spatio-temporal saliency maps to be used for active object recognition in egocentric videos with a particular emphasis on geometrical cue in case when central-bias hypothesis does not hold. The geometrical cue was trained as a circular or flared not-centered Gaussians using a corpus from a recording with the same camera position on the body. We proposed to combine spatial, temporal and geometrical cues as a linear combination with trained coefficients.

The proposed models were compared to other existing reference saliency models with usual evaluation metrics with regard to simulated human visual attention maps and with a quality metric based on the performances of an object recognition framework. The experiments have shown our saliency models to achieve the best performances in this kind of video content. In the future, such saliency models could be applied to other object recognition paradigms and other datasets where central-bias hypothesis does not hold.

4.2 Contribution to top-down saliency prediction

In addition to our bottom-up contribution we propose a new top down probabilistic saliency model for egocentric video content. It aims to predict top-down visual attention maps focused on manipulated objects, that are then used for psycho-visual weighting of features in the problem of manipulated object recognition. The model is probabilistically defined using both global and local appearance features extracted from automatically segmented arm areas and objects. A psycho-visual experiment has been conducted in a guided framework that compares our proposal and other popular state-of-the-art models with respect to human gaze fixations. The obtained results show that our approach outperforms several popular bottom-up saliency approaches in a well-known egocentric dataset. Furthermore, an additional task-driven assessment for object recognition in egocentric video reveals that the proposed method improves the performance of several state-of-the-art techniques for object detection.

4.2.1 Introduction

As presented in chapter 3, two types of attention are commonly distinguished in the literature: bottom-up or stimulus-driven and top-down attention or goal-driven. [Carrasco 11, Pinto 13].

Recent works build class-agnostic object detectors to generate candidate salient bounding-boxes which are then labeled by later class-specific object classifiers [Erhan 14, Shen 14]. As stated in section 3.2.2 the current state-of the art in computer vision allows detection of some categories of objects with a high confidence. A variety of face or skin detectors have been proposed since the last two decades [Jones 99, Viola 04]. Hence, when modeling a top-down attention in a specific visual search task, we can use such "easily recognizable" semantic elements that are relevant to the specific task of the observer and may help to identify the real areas/objects of interest.

In this contribution we propose to use domain specific knowledge to predict top-down visual attention in the task of recognizing manipulated objects in egocentric video content. In particular, our "recognisable elements" that are relevant to the task, are the arms and hands of the user wearing the camera and performing the action. Their quantized poses with regard to different elementary components of a complex action such as object manipulation will help in the definition of the area where the attention of the observer searching for manipulated objects will be directed. We evaluate our model from two points of view: i) prediction strength of gaze fixations of subjects observing the content with the goal of recognition of a manipulated object, and ii) performance in the target object recognition by a machine learning approach.

The rest of this section is organized as follows: we begin by presenting our approach to generate top-down visual saliency maps in part 4.2.2. Section 4.2.3 describes the different experimental set-ups and provides the evaluation of the results. Finally section 4.2.6 draws main conclusions of this work and introduces research perspectives.

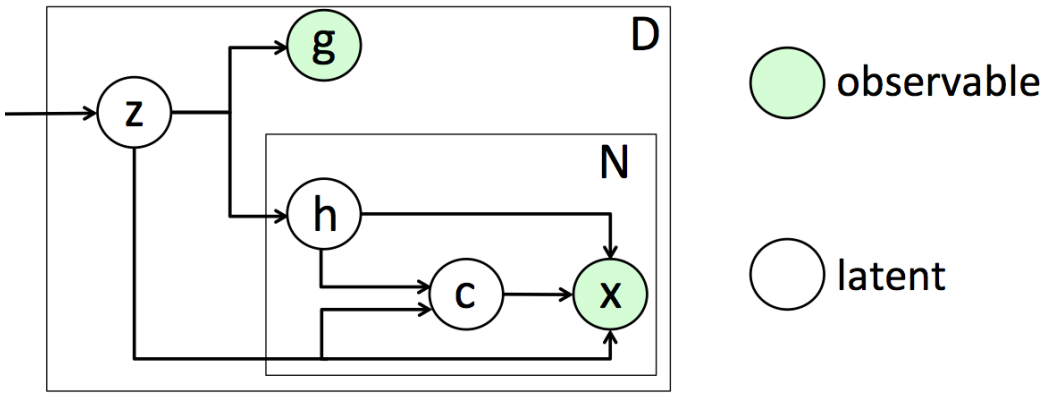


Figure 4.11 – Graphical model of our approach Top-down visual attention modelling with manipulated objects. Nodes represent random variables, edges show dependencies among variables, and boxes refer to different instances of the same variable. Latent variables (transparent background): z set of global arms configurations, h set or arm labels, c set of hand centre positions. Observable variables (shaded-green background): g global features, x spatial locations. D and N refer respectively to the number of training images and number of pixels in a frame.

4.2.2 Goal-oriented top-down visual attention model

In this section we define a model of visual attention prediction in the task of manipulated object recognition. Our model relies on the detection and segmentation of some objects, considered as references, that help to locate the real areas of interest in a scene, namely the objects being manipulated.

We propose to build our model as a combination of two distinct sets of features: global and local. The former describes the geometric configuration of the segmented arms, which are clustered into a pre-defined set of states/configurations. This global information is used to select one of the components in a mixture model. The second set, concerning the local features, is then modeled using the particular distributions corresponding to the selected global component.

4.2.2.1 Defining global and local features

The features we propose are based on the geometry of arms in the camera view field, which is correlated with manipulated object size and position. Each arm, from elbow to the hand extremity, is approximated by an elliptic region in the image plane. Hence an ellipse is first fitted to each segmented arm area and, then, several global features are defined, namely:

- *Relative location of hands*: Two features are extracted that encode the relative location of one hand with respect to the other (see figure 4.12a). For that end, taking the left hand centre as the origin of coordinates, the vector that joins the origin and the right hand is represented by means of its magnitude ρ_{Rel} and phase φ_{Rel} . Magnitude and phase are strong indicators of the objects width and holding pose, respectively.

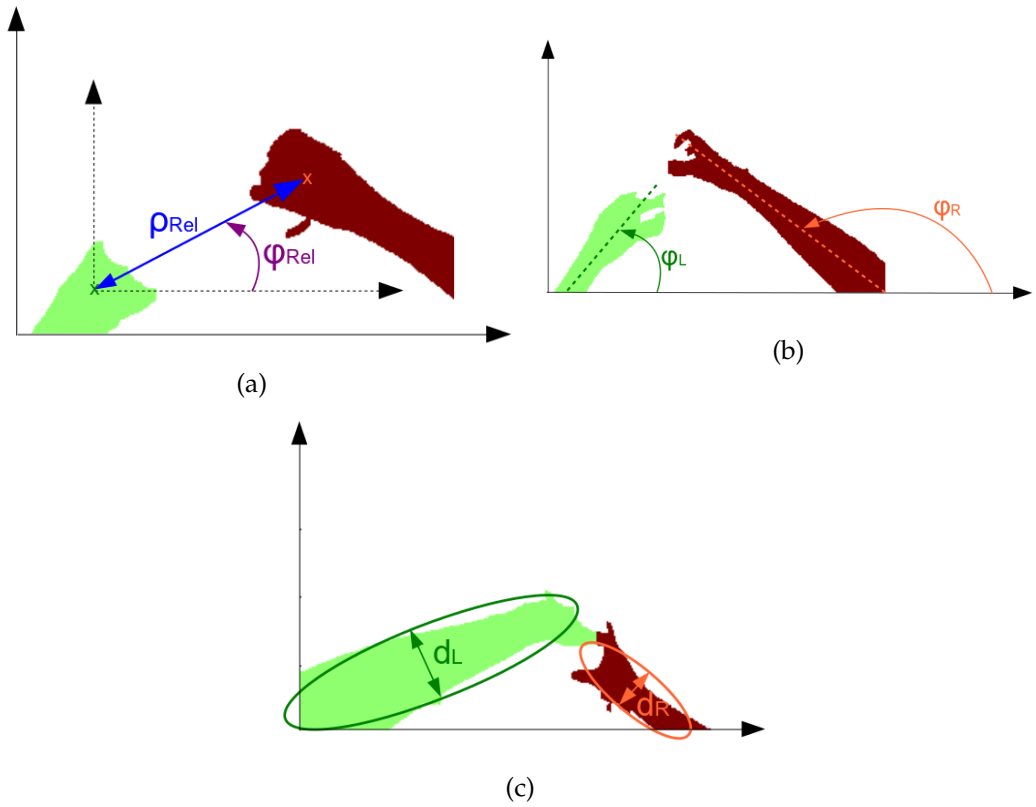


Figure 4.12 – Illustrations of the 6 global features. 4.12a: *Relative location of hands*, 4.12b: *Left arm orientation*, 4.12c: *Left arm depth* and *Right arm depth with regard to the camera*.

- *Left arm orientation* and *Right arm orientation*: As illustrated on figure 4.12b the orientation of each arm (φ_L and φ_R) is defined by the angle between principle axis of ellipse and Y-axis in image plane. The arms are mostly oriented depending on the objects being manipulated, e.g.: holding a cup or pouring something (milk, juice, ...) present usually distinguishable arms orientations.

- *Left arm depth* and *Right arm depth with regard to the camera*: an object size is likely to be correlated with the "depth" of the arms, i.e. a measure of its closeness to the camera. In this work, the body-worn cameras do not provide a real depth information. A trivial approximate of the "depth" of an arm, is the minor axis length d_L and d_R of the fitted ellipse (see figure 4.12c).

A vector $\mathbf{g} = (\rho_{Rel}, \varphi_{Rel}, \varphi_L, \varphi_R, d_L, d_R)$ containing these six geometrical features is computed for each image in the training set, and then clustered into K global appearance models using k-means algorithm. It is worth noting that a Z-score normalization has been performed over the data, in order to prevent outweighing features with large range over attributes with small ones [Al 06]. Figure 4.13 illustrates results in case of 8 clusters in our training dataset. The difference between the global appearance states (a) - (h) is easily noticeable.

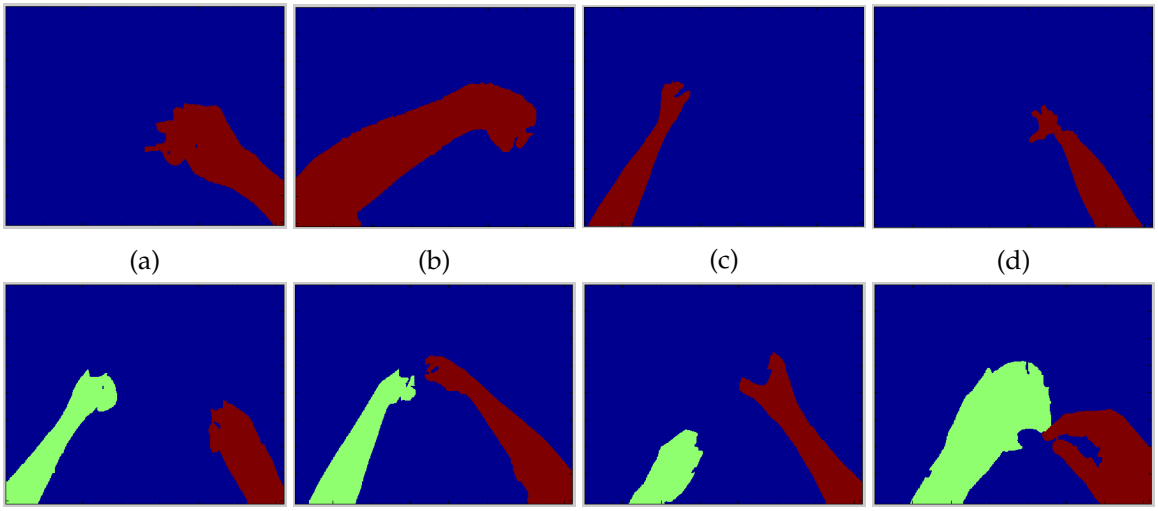


Figure 4.13 – Representation of the arm segmentations closest to the centre of 8 global appearance model clusters. Each cluster is represented by the sample that is closest to the cluster centre.

Furthermore, we consider some "local" features that help to compute a saliency distribution given by the global arm configuration of a frame. These features are the coordinates of hand centres $\mathbf{c}$, the hand indicator h (left or right), and the candidate pixels $\mathbf{x}$ around the hand to belong to the object being manipulated.

4.2.2.2 A Probabilistic Model for Top-down Visual Attention Prediction

As a human observer would be attracted by hand-manipulated objects, we consider the joint locations of arms/hands and objects as predictors of top-down visual attention. Hence our probabilistic model for top-down visual attention incorporates distributions of both global and local features presented in the previous section. The graphical model of our approach is shown in Fig. 4.11. Based on this, given a corpus of D training images our objective is, for each image d, to learn the process that chooses a set of N salient spatial locations $\mathbf{x}$.

To do so, the generative process first randomly picks a global arm model $\mathbf{z}_k$ from the K candidates. K corresponds to the number of clusters as defined in section 4.2.2.1, and remains an open parameter in our model. Then, depending on the selected global model $\mathbf{z}_k$, global ($\mathbf{g}$) and local ($\mathbf{x}, \mathbf{c}, h$) features are drawn from the particular conditional distributions $p(\mathbf{g}|\mathbf{z}_k)$ and $p(\mathbf{x}, \mathbf{c}, h|\mathbf{z}_k)$, respectively. Here, h is an index variable with two possible values $h = 0, 1$ for left and right hands, respectively.

In the following paragraphs we first introduce the distributions modeling both global and local features, then integrate these distributions to build the generative saliency model.

Distributions of Global Features: We define the conditional distribution that models the global features given the component z_k with a Gaussian pdf $p(\mathbf{g}|z_k) = \mathcal{N}(\mathbf{g}; \mu_k^g, \Sigma_k^g)$, with mean vector μ_k^g and covariance matrix Σ_k^g.

Distributions of Local Features: Concerning the local features, for each elementary arms model z_k, we draw N points at spatial locations $\mathbf{x}$ considered as salient. For that end, we start by picking a hand (left or right) following the distribution $p(h|z_k)$. Next, once the hand is chosen, we randomly locate its centre by drawing its coordinates $\mathbf{c}$ using the distribution $p(\mathbf{c}|h, z_k)$. Finally, we use the conditional distribution $p(\mathbf{x}|h, \mathbf{c}, z_k)$ to randomly choose a spatial location $\mathbf{x}$ that belongs to the object being manipulated. This distribution models the probability of a pixel to belong to the object being manipulated given the current geometric configuration of arms and hands.

Putting everything together and marginalizing over the variable h, we can expand the distribution involving the *local features*:

$$p(\mathbf{x}, \mathbf{c}, h|z_k) = \sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k) \tag{4.4}$$

Now, we can define the particular conditional distributions that model each variable:

1. The selected hand is given by a discrete distribution $p(h_j|z_k) = \alpha_{jk}$, with $\sum_{j=0}^{1} \alpha_{jk} = 1$.

2. The hand centre $\mathbf{c}$ follows a Gaussian distribution $p(\mathbf{c}|h_j, z_k) = \mathcal{N}(\mathbf{c}; \mu^c_{jk}, \Sigma^c_{jk})$.

3. The spatial location $\mathbf{x}$ is defined with an experimental discrete distribution: $p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) = \beta_{kji}$, so that $\sum_{i=0}^{L^2} \beta_{kji} = 1$. This distribution is defined over a square 2D box of size LxL centered at $\mathbf{c}$ built by superimposing all accordingly-centered annotated objects from images belonging to the cluster z_k. In Fig. 4.14 we show some empirical examples of this distribution.

Distribution the model: Finally, integrating the distributions of global and local features, the *salience value of a pixel* $\mathbf{x}$ is defined by the following density function, given the parameters $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$:

$$\begin{aligned} S(\mathbf{x}) = p(\mathbf{x}, \mathbf{g}) &= \sum_{k=1}^{K} p(z_k)p(\mathbf{g}|z_k)p(\mathbf{x}, \mathbf{c}, h|z_k) \\ &= \sum_{k=1}^{K} \pi_k p(\mathbf{g}|z_k) \sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k) \end{aligned} \tag{4.5}$$

where $p(z_k) = \pi_k$ is discrete with parameter π_k and stands for the prior distribution of the global arm models (weights of components in the mixture). Let us note that the model in eq.(4.5) allows to compute saliency even in the case where one of the arms is absent by simply considering the corresponding probabilities $p(h = 0|z_k)$ or $p(h = 1|z_k)$ as zero.

To summarize, we have developed a probabilistic model that explains how salient pixels are chosen based on hands/arms configuration and the relative expected location of the object being manipulated within each geometric arrangement.

4.2.2.3 Optimizing the model

From the graph depicted in Fig. 4.11 and equation 4.5, the likelihood of models parameters $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$ can the be defined for the corpus (over all N pixels an D images considered) as:

$$\mathcal{L} = \prod_{d=1}^{D} p(\mathbf{z}_d) p(\mathbf{g}_d|\mathbf{z}_d) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|\mathbf{z}_d) \tag{4.6}$$

If we marginalize eq. 4.6 over the latent arm models we get a definition of the likelihood by means of a mixture of K components:

$$\mathcal{L} = \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \tag{4.7}$$

Taking logarithms and applying the Jensen's inequality (see appendix 8 for further details) one can obtain a lower bound of the log-likelihood:

$$\begin{aligned} \log \mathcal{L} \geq \quad & \textstyle\sum_{d,k}^{D,K} \phi_{dk} \Bigg[\log \Bigg(p(z_k) p(\mathbf{g}_d|z_k) \cdot \\ & \textstyle\prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \Bigg) - \log \phi_{dk} \Bigg] \end{aligned} \tag{4.8}$$

where we have introduced a new variable $\phi_{dk} = p(z_k|\mathbf{g}_d, \mathbf{x})$ which stands for the posterior distribution of the arms model given the observed variables and obeys $\sum_k \phi_{dk} = 1$.

In addition, we can also lower-bound the term of log-likelihood related to the local features (defined by $\mathcal{L}_{local}$) if we apply again the Jensen's inequality:

$$\mathcal{L}_{local} \geq \sum_{dki} \phi_{dk} \sum_{j} \gamma_{dkij} \left[\log p(h_j|z_k) p(\mathbf{c}|h_j, z_k) p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) - \log \gamma_{dkij} \right] \tag{4.9}$$

where $\gamma_{dkij} = p(h_j|z_k, \mathbf{c}, \mathbf{x}_i)$ is the posterior distribution of the selected hand once the global model, the center and the spatial location are known. For more details, the reader is referred to the appendix 8.

Inference: We aim to learn the set of optimal model parameters $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$ that maximize the log-likelihood. For that end, we have used the Expectation-Maximization (EM). Due to the length of the algebra to obtain the EM update equations we omit it in this paragraph. All the details are given in appendix 8.

In the *E-Step*, the algorithm computes the expected values of the posterior distributions ϕ_{dk}, γ_{dkij}:

$$\phi_{dk} \propto p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(x_i, \mathbf{c}_i, h_i|z_k) \tag{4.10}$$

$$\gamma_{dkij} \propto p(h_j|z_k) p(\mathbf{c}|h_j, z_k) p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) \tag{4.11}$$

In the *M-Step*, our algorithm updates the values of the model parameters:

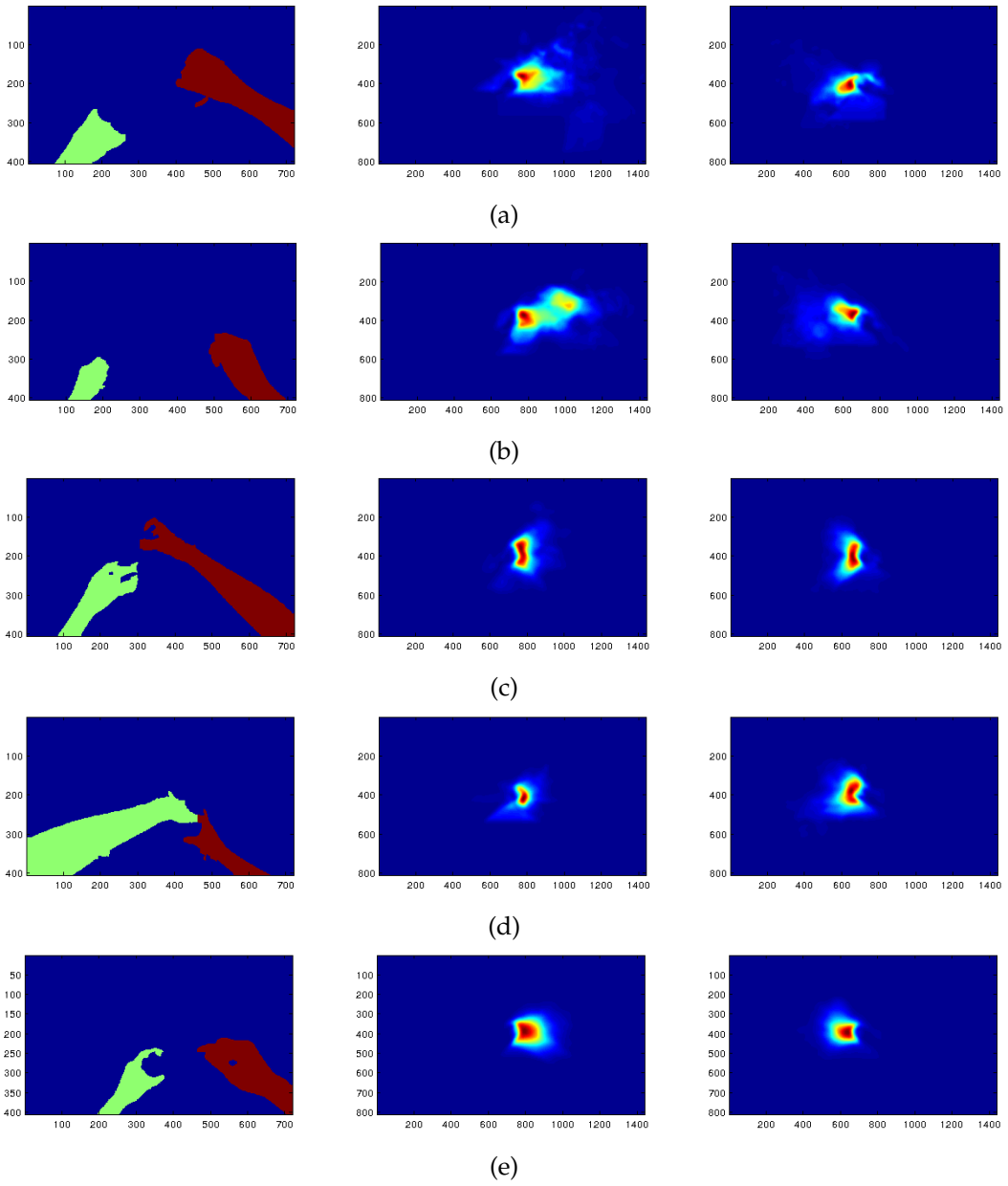


Figure 4.14 – Five examples of the obtained experimental distributions $p(\mathbf{x}|h, \mathbf{c}, z_k)$. Left column: arm segmentation closest to cluster, Middle column: left hand distribution, Right column: right hand distribution.

$$\pi_k = \frac{1}{D} \sum_d \phi_{dk} \tag{4.12}$$

$$\mu_k^g \propto \sum_d \phi_{dk} \mathbf{g}_d \tag{4.13}$$

$$\Sigma_k^g \propto \sum_d \phi_{dk} (\mathbf{g}_d - \mu_k^g)(\mathbf{g}_d - \mu_k^g)^T \tag{4.14}$$

$$\alpha_{jk} \propto \sum_{di} \phi_{dk} \gamma_{dkij} \tag{4.15}$$

$$\mu_{jk}^c \propto \sum_d \phi_{dk} c_{dj} \sum_i \gamma_{dkij} \tag{4.16}$$

$$\Sigma_{jk}^c \propto \sum_d \phi_{dk} (c_{dj} - \mu_{jk}^c)(c_{dj} - \mu_{jk}^c)^T \sum_i \gamma_{dkij} \tag{4.17}$$

Building Saliency Maps: Once the optimal parameters have been learned, we can build a saliency map by measuring the saliency of every pixel location. For that end, the *saliency value of a pixel* $S(\mathbf{x})$ can be defined as its likelihood over the proposed generative model for saliency $S(\mathbf{x}) = p(\mathbf{x}_i, \mathbf{g}|\theta)$

4.2.3 Experimental setup

In this section we present the dataset and provide a whole description of the different experimental set-ups for the comparison of our probabilistic top-down saliency model against other saliency approaches. We also assess its contribution regarding manipulated object recognition performances.

4.2.3.1 Dataset description

In this study we assessed the performances of our top-down model on the GTEA dataset, first introduced in [Fathi 11b]. It is a publicly available database of egocentric videos of 4 subjects performing 7 types of instrumental activities of daily living. The segmentations of arms and objects of interest are provided for 17 videos. The frames were annotated with the objects of interest but we manually extended this annotation by drawing bounding boxes on them. The bounding boxes provide the "ground truth" results that could be reached with an "ideal" rectangular salient area. We did not use the setup proposed in [Fathi 11b], where the authors used videos from 3 subjects to train their system and the last one for evaluation, since the arm segmentations provided with the dataset do not cover all videos from Fathi's setup. Instead we have split the dataset into a training and test set of videos in such a manner as to even the number of samples of each object category in both sets.

For a better understanding, Table 4.5 contains the list of videos belonging to the training and test sets, Figure 4.15 shows the number of occurrences of each category in both sets. Let us note that this set-up can be consireded more challenging than the one presented in [Fathi 11b] since there is less training data and more test data. Furthermore we would also like to explain that, although videos from the same user are contained in both training and test datasets, it does not simplify the recognition task with respect to the original set-up as, in practice, both the scenario and manipulated objects are the same for every user in the dataset.

4.2.3.2 Selected visual saliency models for comparison

The following saliency prediction models were selected for comparison due to their popularity or particular suitability to egocentric video.

- The well-known reference model developed by Itti [Itti 98]. We will denote it as "ITTI" for the rest of this study.
- The graph-based visual saliency model developed by Harel [Harel 07]. It will now be referred to by the acronym "GBVS".

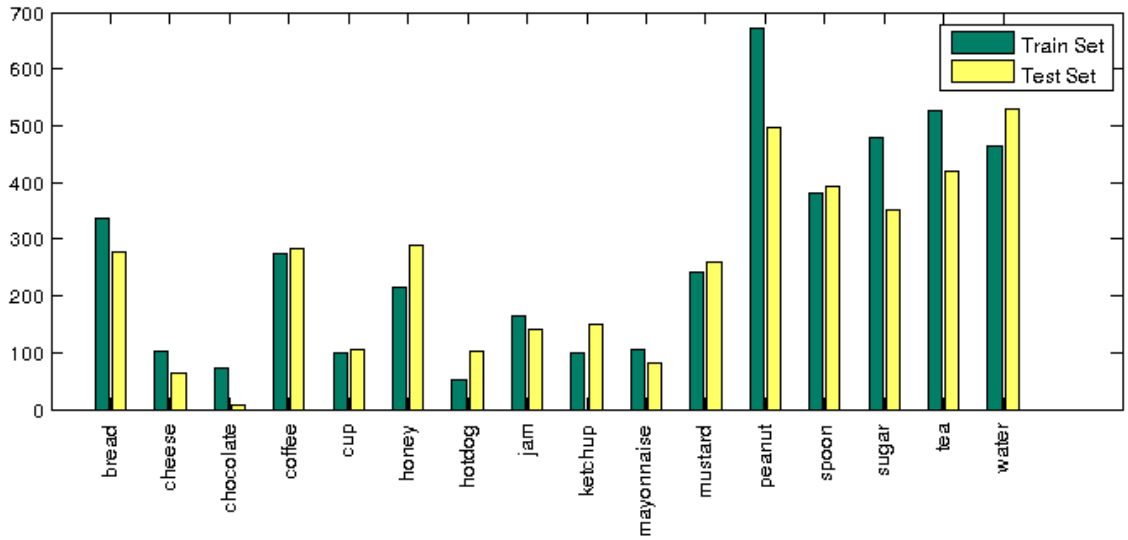


Figure 4.15 – Occurences of each class in our Train and Test sets. The dataset has been split by videos so that the number of samples of each category in both sets is closest.

Training Set	Test Set
S3_Hotdog_C1	S3_Tea_C1
S1_Cheese_C1	S2_Tea_C1
S2_Peanut_C1	S2_Cheese_C1
S2_Coffee_C1	S2_Pealate_C1
S3_Coffee_C1	S1_Coffee_C1
S1_Tea_C1	S1_Hotdog_C1
S1_Pealate_C1	S1_CofHoney_C1
S3_Peanut_C1	S1_Peanut_C1
	S2_Hotdog_C1

Table 4.5 – List of videos in Training and Test sets.

- The spatio-temporal-geometric model presented in [Boujut 12a] since it has been specifically developed for saliency extraction in egocentric videos and presents the state-of-the art in saliency-based object recognition in this content [González Díaz 13]. This model will be referred as "STC" as previously introduced in section 4.1.1.
- Visual Attention maps built on gaze fixations by reference Wooding's method [Wooding 02]: the fovea projection for each fixation is modelled with a Gaussian of two visual degrees spread and resulting multi-Gaussian surface is normalized.

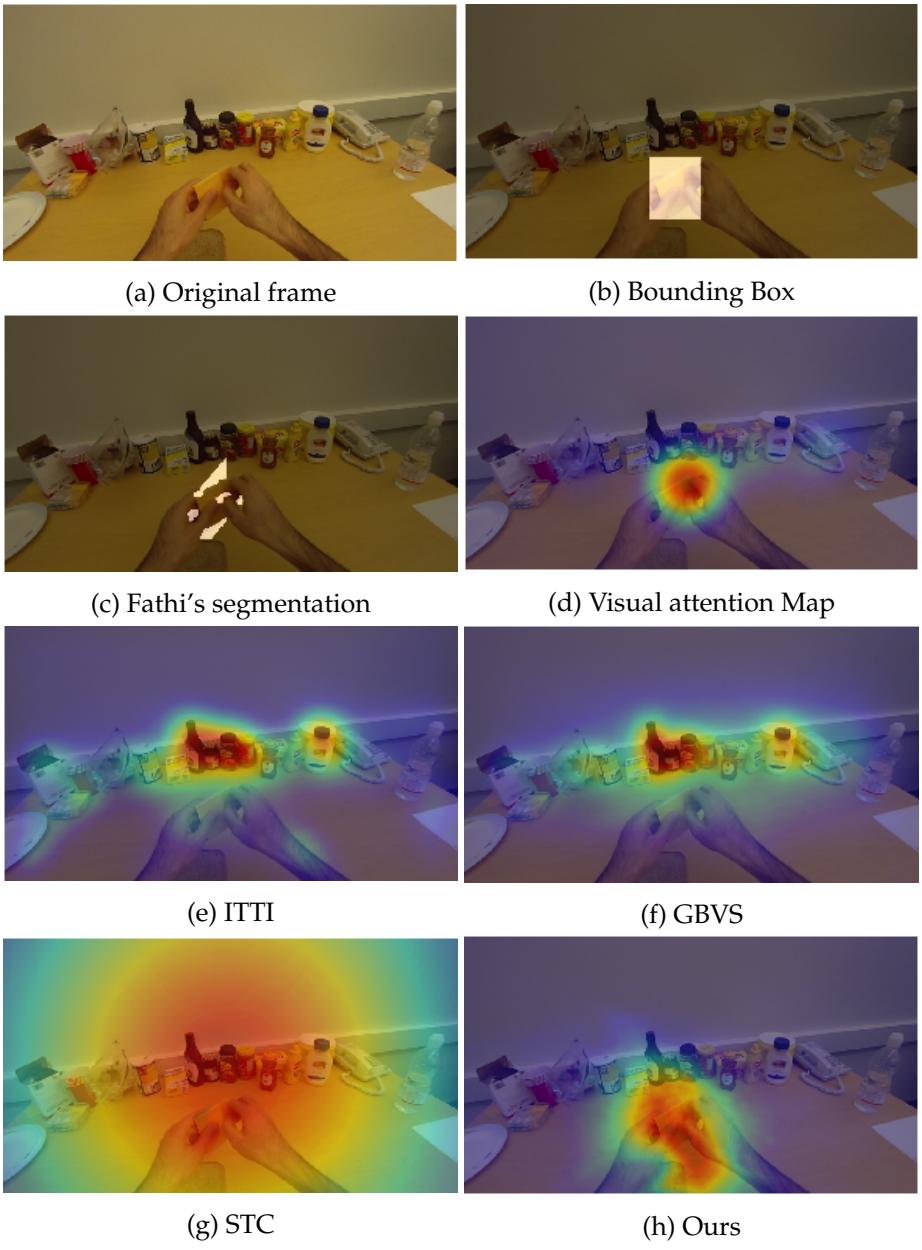
Figure 4.16 contains computed saliency maps for a randomly selected frame (a). We also display the manually annotated bounding box of the manipulated object (b), as well as the automatically extracted segmentation mask (c).

4.2.4 Psycho-visual evaluation of proposed saliency model

In this section we assess the capacity of our top-down model to predict human visual attention in the task-guided psycho-visual experiment. The saliency models presented in section 4.2.3.2 were also assessed for the sake of comparison. The psycho-visual experiment was designed for recording gaze fixations of subjects who observed the egocentric video with the task of recognition of manipulated objects. For this experiment 31 participants have been gathered, 10 women and 21 men. They were given a written instruction to look specifically at the manipulated object in videos. Each video was watched by at least 15 subjects. The gaze positions have been recorded with a HS-VET 250Hz Cambridge Research Systems Ltd eye-tracker. The experiment conditions and the experiment room were compliant with the recommendation ITU-R BT.500-11 [ITU 02]. Videos were displayed on a 23 inches LCD monitor with a native resolution of 960×540 pixels. To avoid image distortions, videos were not resized to screen resolution but instead a grey frame was inserted around the displayed video. In order to avoid the visual fatigue, the duration of observation was not longer than 15 minutes for each subject.

Automatically predicted saliency maps can be compared to human gaze fixations with the help of dedicated metrics. From [Riche 13] and previous work [LeMeur 12], we retained the NSS (for more information, see 3.4.1.3)

We measured the similarity of recorded eye fixations from the experiment with automatically generated saliency maps from our top-down probabilistic model and the ones presented in section 4.2.3.2. In total 8244 frames were compared for each saliency model and the final mean scores with standard deviations are presented in Table 4.6. As shown in the table, our proposed top-down probabilistic model corresponds better to real human eye fixations than the other state-of-the-art saliency models. Since the standard deviation are high, we computed the p-values to back up the hypothesis that the NSS mean using our top down approach is significantly higher than with the other attention prediction models. At the 5% significance level, the data do provide sufficient evidence to conclude that the mean NSS score using our top-down saliency is greater than the mean obtained using other saliency models. It is however important to underline that the GVBS and ITTI models are bottom-up and were not designed for a task of recognition of specific objects of interest.



(a) Original frame (b) Bounding Box

(c) Fathi's segmentation (d) Visual attention Map

(e) ITTI (f) GBVS

(g) STC (h) Ours

Figure 4.16 – Saliency models selected for comparison.

4.2.5 Object recognition performances

The ultimate goal of developing a model of top-down visual saliency is in the task of manipulated object recognition. Hence, we first discuss the object recognition approach with saliency-based psycho-visual weighting of features. This approach, combined with the proposed saliency model, is then compared to other state of the art paradigms for object recognition. We also benchmark it with other saliency models presented in section 4.2.3.2.

	ITTI	GBVS	STC	OURS
mean NSS score	1.05 ± 0.7269	1.29 ± 0.6551	1.52 ± 0.2490	$\mathbf{2.28} \pm 1.2226$

Table 4.6 – NSS mean scores (with standard deviations) between human fixation points and different saliency map models. Our model outperforms the others

4.2.5.1 Saliency-based object recognition approach

In this study we used the same saliency-based object recognition method presented previously in this chapter in 4.1.2.4.

Since the saliency depends on the segmentation of the arms it is possible to find cases where arms do appear on the image or are not detected by the segmentation algorithm (this has happened only in 720 cases, meaning around 4.1% of all the segmentations provided by Fathi in the dataset). In these cases our model obviously does not provide a saliency map and it is up to the user to decide which saliency model to use. The models in section 4.2.3.2 constitute valid alternatives among which the STC ([Boujut 12a]) stands out as it has been specifically developed for saliency extraction in egocentric videos. In this work however, in order to rely solemnly on our model during the computation of performances, no other saliency model was computed to replace cases where arms are not detected. Instead we chose to build non-weighted signatures as in the original BoVW framework.

For the computation of BoVW, we use a dictionary size of 4000 visual words. Once each image is represented by its weighted histogram of visual words, an SVM classifier [Cortes 95] is used with χ^2 kernel. Posterior probabilistic estimates for the occurrence of the object of class C in the frame t are finally obtained using Platt's approximation [Platt 99].

4.2.5.2 Influence of the number of clusters in the global appearance model

The number of clusters K introduced in section 4.2.2.1 is an open parameter in our model. We have performed an optimization of the target mAP of object recognition in regard to this parameter using the paradigm previously introduced in section 4.2.5.1. Table 4.7 below illustrates the influence of the number of clusters K to the target mAP. Having too few clusters might lead to a lack of information about certain arm models while having too many leads to poorly populated clusters. The case of $K = 1$ is the specific case where we do not consider the information given by global features. We observed that its high generality makes it perform well in most categories of objects. However, some categories of objects with specific shapes ("water", "ketchup", "sugar", ...) are manipulated in certain ways such that removing global features yields a drop in recognition performances.

For the rest of the experiments the saliency model referred as "Ours" corresponds to the methodology presented in section 4.2.2 with $K = 50$ clusters, which has turned out to be the optimal value in our experiments.

	$K = 1$	$K = 20$	$K = 50$	$K = 100$
mAP	0.301	0.316	**0.353**	0.342

Table 4.7 – Validation of the number of global appearance models K

4.2.5.3 Influence of the arm segmentation performances

As stated previously, this work aims to provide a model for computing top-down semantic saliency maps given the arm segmentations. Hence the performance of our approach is deeply linked to the quality of the arm segmentation. In this part we aim to study how much segmentation errors could alter the performance of our proposed model. It is possible, based on the data provided in this dataset, to alter the given arm segmentations by applying varyingly important transformations to the previously segmented arms (e.g. homographies). However in order to truly degrade segmentation performances, we chose to implement a genuine hands segmentation framework and train it with different amount of training data.

Detection of hands/arms in egocentric videos has already been the core of several recent studies ([Li 13b, Lee 14, Li 13a, Betancourt 14]). In this work, we retained the framework of [Li 13b] which has shown to provide good performances in similar contents. It is based on training modls for hand (arm) pixels with a training set of patches. Then a binary classification of pixel is performed. This segmentation paradigm was pioneer in the domain since it was the first to propose a model adapting to different illumination conditions, which proved to be essential in egocentric videos where lighting conditions vary often. Figure 4.17 shows some examples of how the segmentation gets affected by varying the number of training data.

In order to measure the segmentation performances 327 images were randomly chosen among the whole dataset and were manually segmented. We therefore compared the similarity between automatic segmentation model and manually annotated data using the Jaccard's similarity coefficient:

$$J(S_m, S_a) = \frac{|S_m \cap S_a|}{|S_m \cup S_a|} \tag{4.19}$$

where S_m and S_a respectively stand for manual and automatic segmentation. Figure 4.18 shows the average similarity between Li's segmentation ([Li 13b]) and the 327 manual segmentations based on the amount of training data. We can see that the segmentation similarity with the ground truth grows with the amount of training data until convergence. The rise of performance is more pronounced for small numbers of training samples, and good similarity scores are rapidly reached (65% for 16 training samples). For 78 training samples, Li's segmentation obtains a similarity score equal to Fathi's, which gets even slightly outperformed for higher numbers of training samples until stabilization around a score of 71.5%. The standard deviation is however almost twice as small as the one obtained with Fathi's segmentation. The rationale behind is that Fathi's segmentation does not always detect arms leading to a Jaccard's coefficient of 0 but, when it does, provides segmentations that are very close to the ground truth.

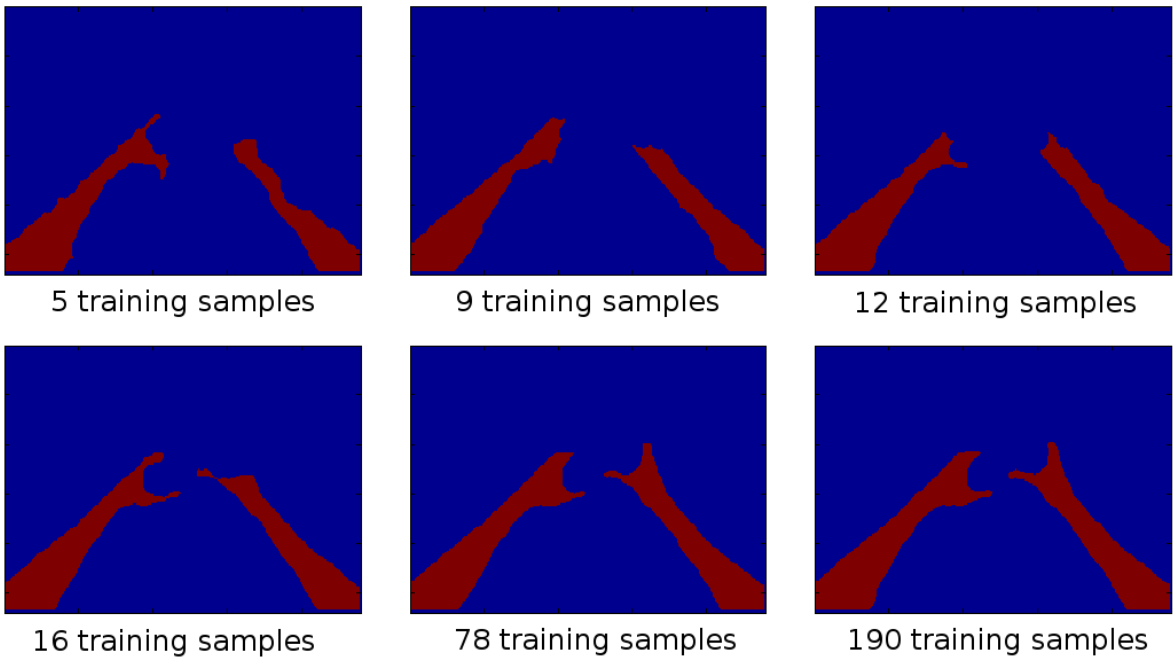


Figure 4.17 – Illustration of arm segmentation outputs with Li's model ([Li 13b]) for different amount of training data

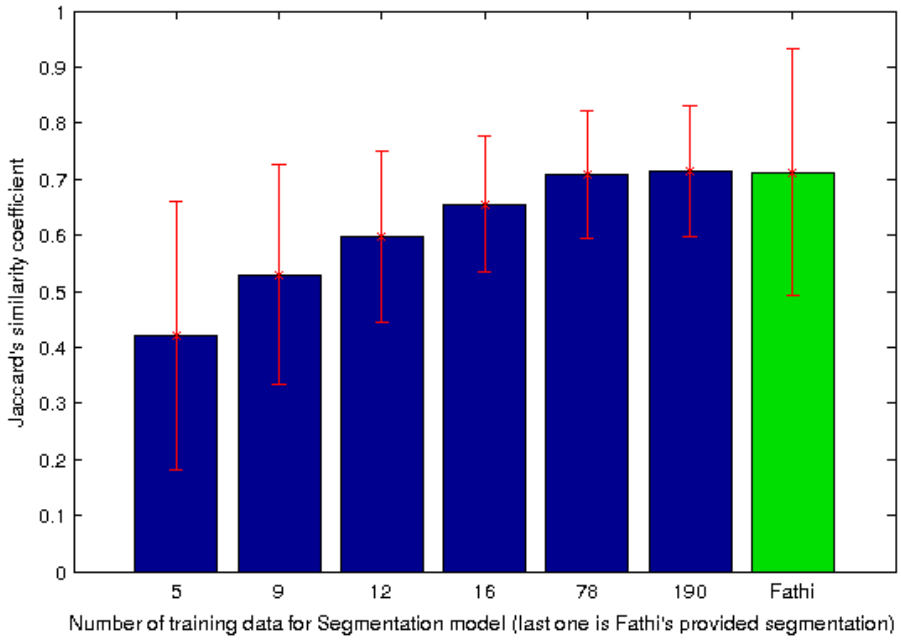


Figure 4.18 – Average similarity between Li's segmentation ([Li 13b]) and the 327 manual segmentations based on the amount of training data. The last column is the similarity with Fathi's provided segmentation.

Number of training samples for Arm segmentation models	5	9	12	16	78	190
mAP	0.309	0.299	0.344	0.347	0.352	0.356

Table 4.8 – Object recognition performances for different number of data used to train the arm segmentation models

In Table 4.8 we present object recognition performances as mean Average Precision scores based on the number of data used to train the arm segmentation models. As expected, there is a significant drop of performance for low number of training data (and hence poor arm segmentation). A gap of more than 5% in mAP is noticeable between the lowest and highest object recognition scores. Two observations can be pointed out from these values however:

- As for the similarity scores in Figure 4.18, the variation of performance is not linear. We notice indeed that performances stay at their lowest point for segmentation similarity scores below 55% but abruptly raise and even start reaching convergence when getting closer to 60%.
- Even for a very low number of training data leading to notably poor arm segmentation, our top-down model, coupled with the object recognition paradigm presented in section 4.2.5 still achieves higher performances with a simple BoVW framework (mAP of 0.246). This can be explained by the modularity of our model. Indeed, we observed that the K global Arm Models introduced in section 4.2.2.1 adapt to the poor segmentation by creating arm models even for these cases and learning an adequate experimental distribution $p(\mathbf{x}|h = j, \mathbf{c}, z_k)$.

4.2.5.4 Comparing with other object recognition approaches and saliency models

For the sake of comparison, we have compared our approach with a baseline model that implements a BoVW without any saliency maps, using a dense sampling of features on the whole frame. This method is referred as "Simple BoVW" in the experiments. In addition, we have also included in the comparison a "ground truth" model where descriptors were extracted only in manually annotated BB. In this method, referred as "BoVW with BB", we consider the ground truth bounding boxes as "ideal" saliency maps.

Figure 4.19 shows the category detailed and average results for the object recognition. As can be seen from the mAP score (last set of bars), our method outperforms the two famous paradigms for object recognition in this kind of video content: i) it achieves an absolute improvements of 10.7% with respect to the baseline BoVW, and ii) a 8.6% absolute improvement with respect to the Deformable Part-Based Model (DPM) from [Felzenszwalb 10]. In addition, also achieves close performances to the "ideal" case, which was added for the upper bound estimate.

In section 4.2.2 we already raised the question of the need of building saliency maps if the objects have been already segmented. Indeed, segmentation as such cannot be used in

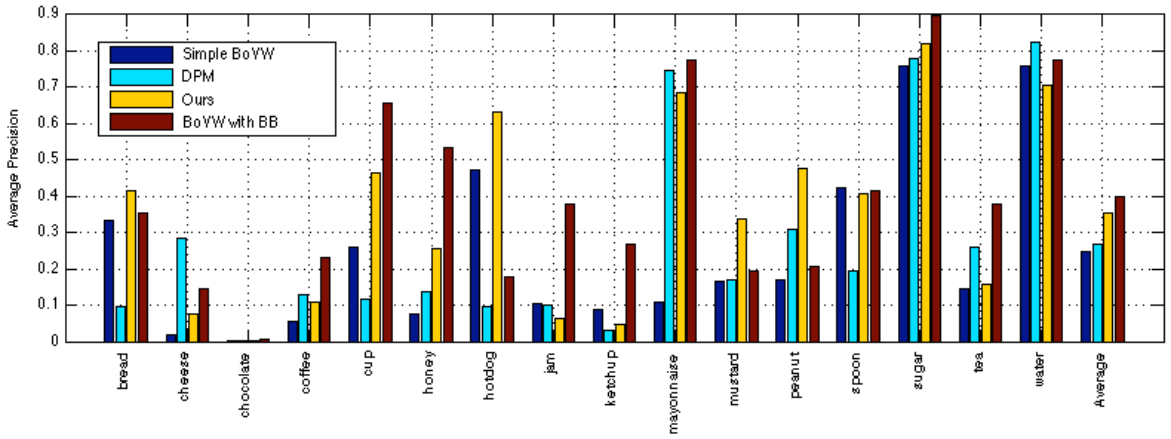


Figure 4.19 – Object recognition performances between different paradigms. The results are given in average precision per category and averaged.

our object recognition paradigm, as segmented objects are often represented by very small and sparse areas (see an example in Figure 4.16c). The extraction of relevant descriptors is thus strongly affected and yields a mAP of only 0.07. Nevertheless, one could object that segmented objects are too restrictive for a comparison to be possible. In this regard, we computed trivial saliency maps as a 2-dimensional fitted Gaussian on the segmented objects, allowing in the process to unify cluttered zones. Such saliency maps gave an object recognition mAP of 0.21, which is still very far from the score of 0.35 achieved by our model.

In their paper, Fathi et al. [Fathi 11b] use a different object recognition method based on the segmented zones. We also computed object recognition accuracy in our test set in the same way it was computed by Fathi. As can be seen in Figure 4.20 the precision obtained by our approach was slightly higher in average that Fathi's. However it is important to note that the comparison is unfair since, as we already mentioned in section 4.2.3.1, we have evaluated our detectors under a more challenging set-up with less training data and more test data. Also an interesting thing to point out is how different both approaches detect better certain categories than other.

We also compare our model with those described in section 4.2.3.2 using the same object recognition approach. Results for per-category and averaged object recognition are displayed in Figure 4.21 in terms of AP. Compared to ITTI and GBVS models, our model performs better for almost all categories. These bottom-up saliency models are stimuli-driven, make use of spatial contrast and were not designed to model a top-down, intentional attention component. The performances of bottom-up STC saliency maps, developed for video were also beaten for almost all categories. This is due to the overestimation by STC of the spread of Gaussian expressing central bias hypothesis on visual attention.

It also achieves slightly better performances than the ones provided by Human Visual Attention maps [Wooding 02]. It is indeed better for some categories since as illustrated in Figure 4.16d, the visual attention maps are perfectly located but sometimes do not cover the objects of interest enough, contrarily to our model (see Figure 4.16h for an example).

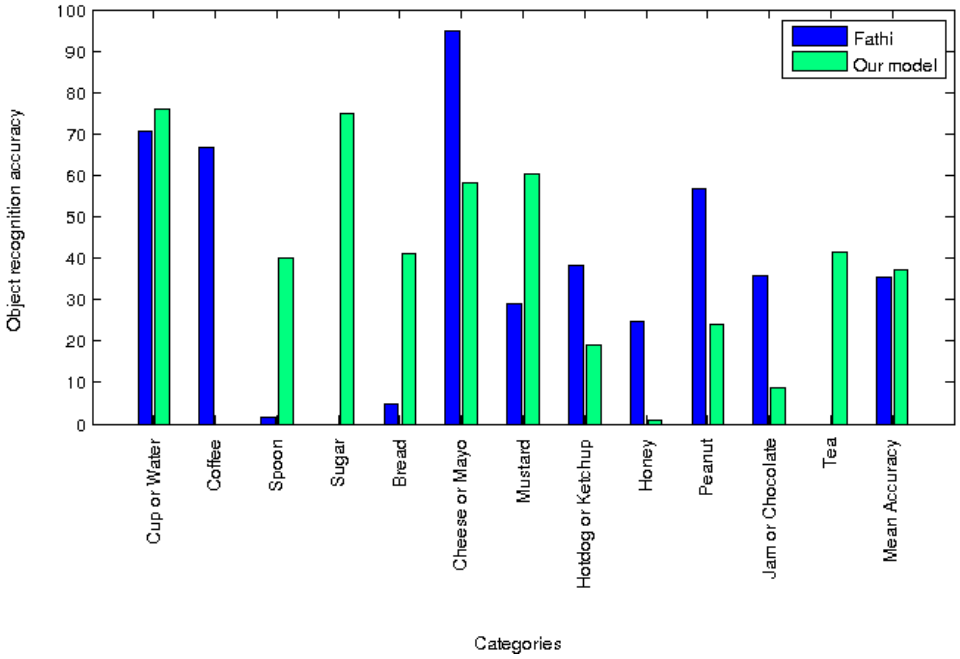


Figure 4.20 – Object recognition accuracy comparison between the model presented in [Fathi 11b] and our approach.

	ITTI/Ours	GBVS/Ours	STC/Ours
p-value	0.0876	0.0118	0.0541

Table 4.9 – p-values between the population consisting in category AP values (mAP) from our method and the ones obtained with AP of each of the other automated Saliency prediction methods (ITTI, GBVS, STC)

On Figure 4.21 we can see it is not necessarily that our top-down model outperforms other saliency methods in each category. We want to find out if the mean value of the population consisting in category AP values (mAP) from our method is significantly different from the mAP obtained with each of the other automated Saliency prediction methods (ITTI, GBVS, STC). Hence we performed Student's t-tests with significance level of 0.10 for comparison and found the null hypothesis to be consistently rejected (p-values provided in table tab:pValuesSaliencies).

4.2.6 Discussion and perspectives

In this second contribution we have proposed a top-down probabilistic visual saliency model for the target task of recognition of manipulated objects in egocentric video. It is based on global and local features and uses domain knowledge, i.e. the fact that the object of interest is manipulated by hands. The model predicts well human attention in a task-driven psycho-visual experiment and shows better performances than several bottom-up models

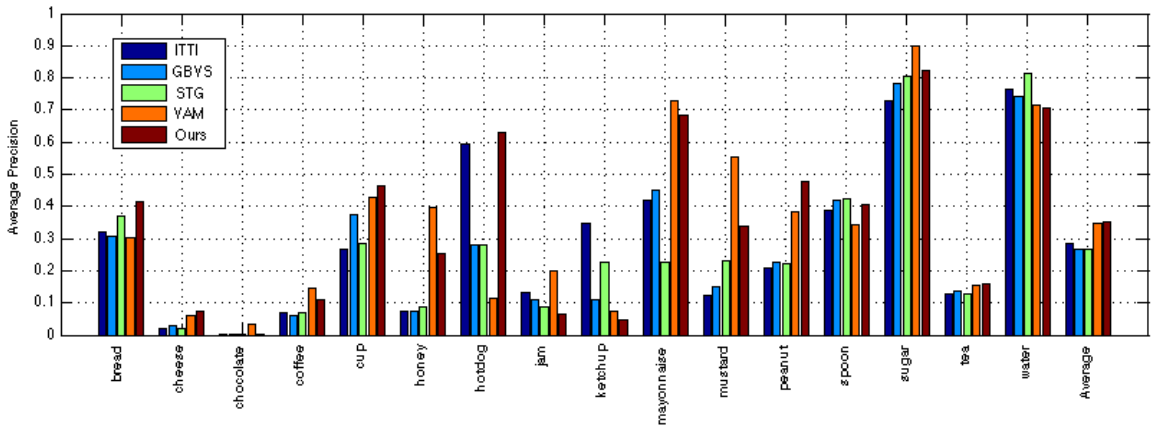


Figure 4.21 – Object recognition performances between different saliency models applied to the saliency weighted BoVW paradigm.The results are given in AP per category and averaged.

widely used in literature, both in terms of comparison with human gaze fixations and target performance in manipulated object recognition task.

Despite the fact that this model has been developed for the specific case of egocentric video content and the task of manipulated object recognition, the idea behind is generic. It is indeed our belief that this model could be extended to other domains of application and not only egocentric videos with detection of arms. The model could be adapted to many scenarios where there exist reference objects, which can be easily recognized, and where the top-down attention is related to them. One interesting example is the aided robotic surgery or the generation of post-surgery video reports. Here, the reference objects are the medical instruments, so that the attention is driven to the close operation field. Further examples are the recognition of e.g. robot-sorted objects on a conveyor belt, carried objects by a crane in a surveillance scenario. Another example, again with egocentric video content, is the real-time detection of objects with wearable glasses for manipulation by neuro-prostheses. Anyway this is a general principle: task-driven visual attention can be easily predicted if we can detect the presence of reference objects for such a task, which in this work where the hands of the user are performing the action.

4.3 Conclusions and discussion

In this chapter we presented two contributions to visual attention modeling in both bottom-up and top-down domains.

In the first contribution we have presented a method for building bottom-up spatio-temporal saliency maps to be used for active object recognition in egocentric videos with a particular emphasis on geometrical cue in the case where a central-bias hypothesis does not hold. We proposed to combine spatial, temporal and geometrical cues as a linear combination with trained coefficients. The experiments have shown promising results as they highlight the necessity of a non-centered geometric saliency cue.

In the second contribution we have proposed a top-down probabilistic visual saliency model for the target task of recognition of manipulated objects in egocentric video. It is based on global and local features and uses domain knowledge, i.e. the fact that the object of interest is manipulated by hands. The model predicts well human attention in a task-driven psycho-visual experiment and shows better performances than several bottom-up models widely used in literature, both in terms of comparison with human gaze fixations and target performance in manipulated object recognition task.

In visual attention modeling we need to use domain knowledge and contextual information. Visual attention is a complex combination of bottom-up, stimuli driven, and top-down, intentional components. In the perspective of the present research, combining of bottom-up and top-down prediction and spatio-temporal evolution of visual saliency in a video scene is envisaged with a target application to object and action recognition.

Now that we presented our contributions for saliency modeling, the next chapter will be dedicated to studying the integration of visual saliency into the challenging task of active object recognition.

Chapter 5

Saliency-based Object Recognition in egocentric videos

5.1 Introduction

As stated in chapter 3, modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered *of interest* or *salient*. Due to the use of saliency maps, the search for objects in images is more focused, thus improving the recognition performance and additionally reducing the computational burden. Even more, saliency methods can be naturally applied to both BoW [Ren 10] and sliding window approaches [Alexe 12, Uijlings 13].

Various authors have shown how extracting visual features in salient areas improves the system performance in several computer vision tasks, such as image retrieval [de Carvalho Soares 12], object recognition [Sharma 12, San Biagio 14], object tracking [Mahadevan 13, Su 14], or action recognition [Vig 12, Mathe 12]. However, although much fundamental work has been done to generate good representations of visual saliency from still images or video content, their application to object recognition has not been yet explored in-depth. Indeed, it is still commonly restricted to a pre-processing stage that filters out non-relevant areas from the process [Ren 10].

In this chapter, therefore, we provide a systematic study of the application of saliency to the challenging task of active object recognition. We aim to model the retina in the HVS by considering biologically-inspired independent foveal and peripheral visual paths. By plugging our contributions in the BoVW paradigm, we investigate how visual attention modeling can be applied to various modules in the processing pipeline. To the best of our knowledge, this is the first in-depth study about the application of visual saliency to object recognition with BoVW approach at all its stages: i) we extend the state-of-the-art on

Saliency-sensitive non-uniform feature sampling in a new *Saliency-sensitive variable-resolution feature space*, ii) we introduce a completely new *Saliency-Sensitive Coding of features* and use the iii) *Saliency-based feature pooling* which has been shown to be efficient in referenced research [González Díaz 13, de Carvalho Soares 12].

The benefits of our approach are multiple: i) the computation of saliency maps is category-independent and a common step for any object detector, ii) compared to sliding window methods, by looking at the salient area we can avoid much of the computational overhead caused by an exhaustive scanning process, iii) our automatic saliency maps not only focus on the object of interest of a scene but usually contain some context around the object, iv)an object recognition method working with saliency maps does not need ground-truth bounding boxes for training, which dramatically reduces the human resources employed in the database annotation. In contrast, a known limitation of the use of saliency is that, as it focuses on the objects/area of interest of the scene, it may prevent systems from detecting non-salient small objects that belong to the background.

In order to assess these benefits, active object recognition in egocentric video content has been selected as our main experimental benchmark. This decision stems from several reasons. First, egocentric video analysis has recently gained a lot of attention due to the emerging end-user applications involving the use of wearable cameras in scenarios such as robotics, telemedicine or life-logging [Karaman 14]. Second, in the particular case of egocentric view, we claim that an action can be effectively defined as a sequence of 'active' objects [González Díaz 13]: objects that are interacted with (manipulated, observed) by the user and constitute the area of interest in the scene. In addition, in egocentric video there is usually a strong differentiation between active and passive objects and, therefore, spatial, temporal and geometric cues can be found which identify the active elements in the scene. For those reasons, other authors have previously applied visual saliency to egocentric video analysis [Fathi 11b, Ren 10, Fathi 12, Ogaki 12].

The remainder of the chapter is organized as follows: Section 5.2 describes in detail our saliency-based approach for object recognition. In section 5.3 an in-depth evaluation is provided that assesses our model under the various scenarios, and compares it to other state-of-the-art approaches. Finally, section 5.4 summarizes our conclusions and gives perspectives.

5.2 A saliency-based approach for active object recognition

In this section we will describe our approach for active object recognition using saliency. As shown in Figure 5.1, we take the BoVW paradigm as our baseline, and propose to improve its spatial precision using saliency maps.

We inform the reader that this section corresponds to a joint work with Dr Ivan Gonzalez Diaz. However it is important to specify that the original idea and implementation of the Variable Resolution Sampling (VSR) and SC techniques are from him.

Our baseline implementation of the BoVW is briefly described as follows: for each frame/image, we extract a set of N local descriptors using a dense grid of overlapped circular patches. Based on several experiments, we have set the radius of the circular patches

to 30px, and the step size between each local patch to 6px. Next, each local patch $n = 1..N$ is described using a 64-dimensional SURF descriptor x_n [Bay 08], which has shown similar performances to the SIFT descriptor [Lowe 04] in our experiments. Each descriptor x_n is then assigned to the most similar word $b_k, k = 1..K$ in a visual vocabulary by following a vector-quantization process. The visual vocabulary B, computed using a k-means algorithm over a large set of descriptors in the training dataset (we use about 1M descriptors), has a size of K visual words. The vector-quantization process allows the generation of image signatures as L1-normalized histograms H of word occurrences. Finally, to detect the presence of a category in the image, we use an SVM with a nonlinear χ^2 kernel, which has shown good performances working with normalized histograms [Sreekanth 10].

In parallel, our system generates a saliency map S of the frame which is used to model two differentiated pathways found in retinal vision: *foveal or central* vision, and *peripheral* vision. It is known that, due to the varying morphology of neurons in the retina, the human eye simultaneously allows for a high-resolution and detailed perception in the visual field associated with the fovea, and a low-resolution one in the peripheral visual field [Wandell 95].

The human perception of a scene is based on information acquired during periods of relative gaze stability known as fixations [Liversedge 11]. For each fixation, a well-defined location of the image corresponds to the fovea location, whereas the rest of the image is associated with the peripheral visual field. Consequently, given the saliency value of an image location, our system models both pathways at various stages of the processing pipeline (denoted with red dotted lines in Fig. 5.1).

In the following sections, we will describe each processing module using saliency. It is worth noting that our objective is to improve system performance by modeling differentiated pathways for foveal and peripheral vision, while keeping the computational burden of the final solution as bounded as possible. Hence, an important requirement is that the enhancement in performance is not achieved at the expense of a dramatic increase in the computational time.

5.2.1 SP

This section describes the first and most trivial way to combine saliency with the BoVW paradigm. In the traditional BoVW approach see 2.3.1.1, the image signature H is the statistical distribution of the image descriptors according to the visual codebook. This is made by first assigning each local descriptor to a visual word in the vocabulary, and then computing a histogram of word occurrences by counting the times that a visual word appears in an image.

In our *Saliency-based Pooling*, we use saliency to weight the selected features, giving place to a sort of soft-assignment based on saliency maps. In particular, the contribution of each image descriptor is defined by the weight s_n in eq. (5.3). In other words, descriptors over salient areas will get more weight in the image signature than descriptors over non-salient

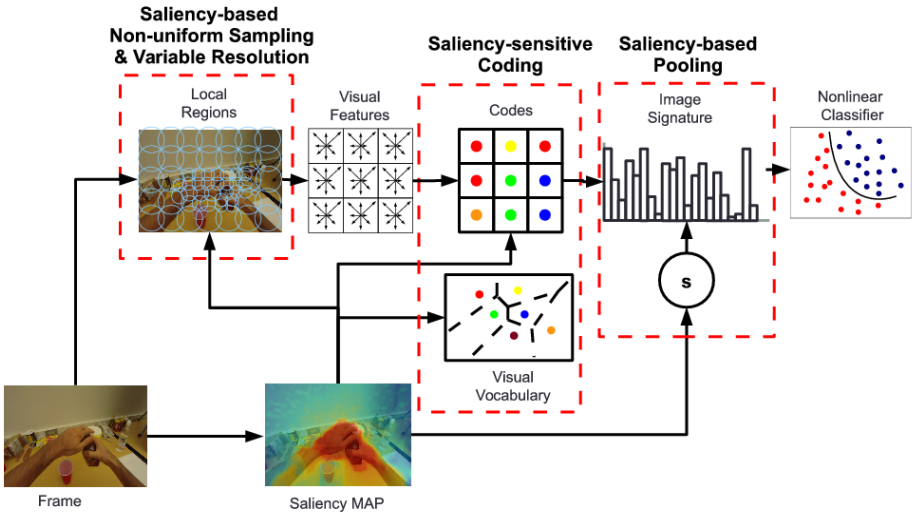


Figure 5.1 – A general view of the processing pipeline for saliency-based object recognition in first-person camera videos, where the three modules that incorporate saliency are surrounded with red dotted lines.

areas. Therefore, the image signature can be computed as follows:

$$H_k = \sum_{n=1}^{N} s_n \alpha_{nk} \tag{5.1}$$

where H_k represents the k-th bin of a histogram, and α_{nk} is an index variable so that $\alpha_{nk} = 1$ for the visual word in the vocabulary associated with the n-th descriptor in the image and $\alpha_{nk} = 0$ for the rest.

Finally, the histogram H is L1-normalized. This method of saliency weighting is similar to the spatial weighting proposed in [Marszałek 06] but, in our case, the weights are not learned from data as, in contrast, are directly derived from saliency, therefore being category-independent.

Furthermore, an extension of the basic saliency-pooling has been explored in [de Carvalho Soares 12], where the authors considered two independent signatures, foreground and background ones, which were defined using a soft fuzzy approach based on saliency. This method can be directly plugged into our perceptual approach modeling our two pathways in retinal vision. Hence, the image signature would be a concatenation of two histograms $[H_f, H_p]$:

$$H_f = \sum_{n=1}^{N} s_n \alpha_{nk}; \quad H_p = \sum_{n=1}^{N} (1 - s_n) \alpha_{nk} \tag{5.2}$$

where H_f stands for the foveal channel, while H_p models the peripheral one. If we keep the vocabulary length K fixed, it will produce image signatures of length $2K$, with a consequent increase in the computational complexity. Alternatively, if we divide the vocabulary

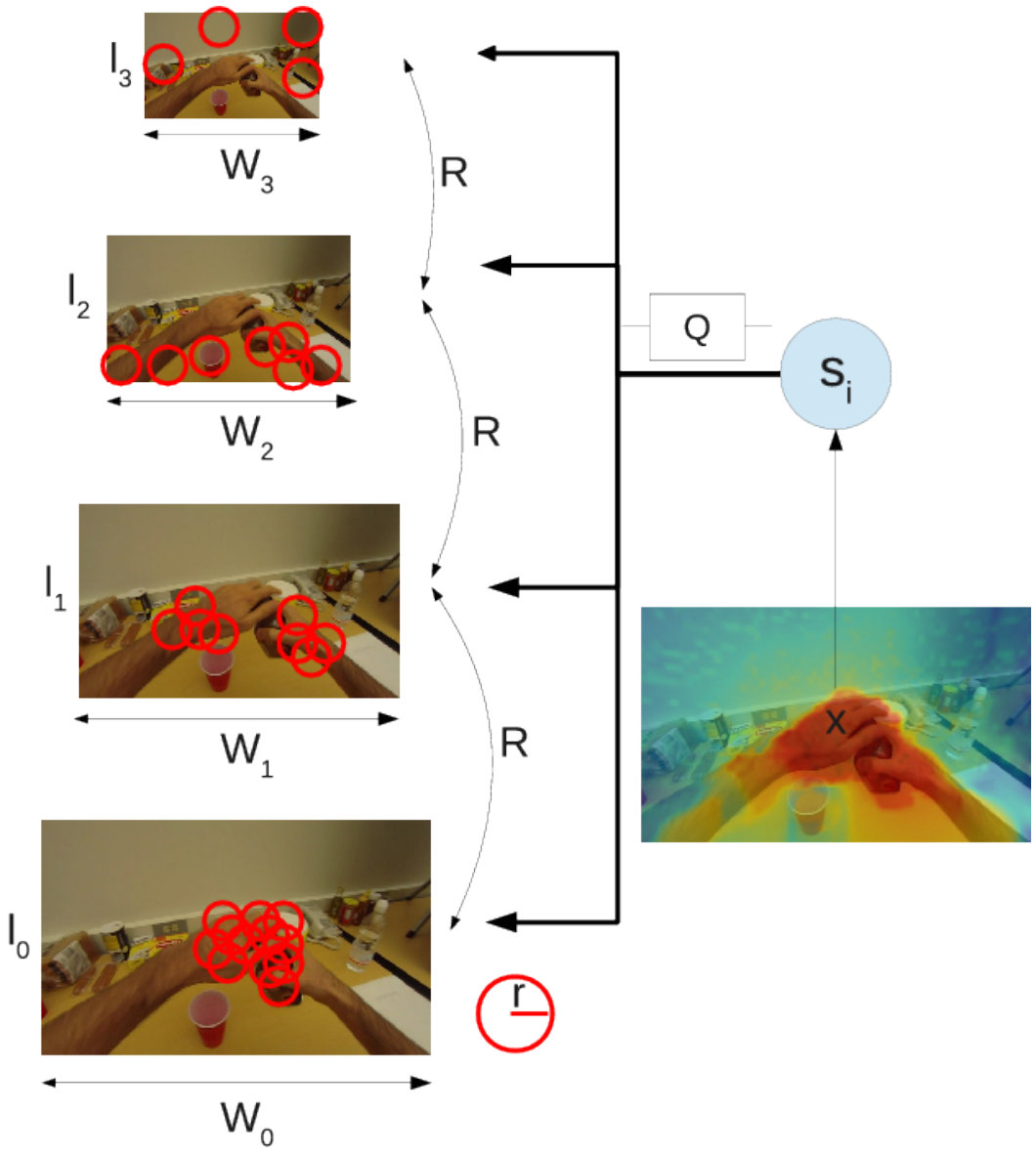


Figure 5.2 – Details of the module "Saliency-based Non-uniform Sampling and Variable Spatial Resolution", where fixed-size circular patches are sampled over an image pyramid based on saliency values.

length by two and keep the computational complexity constant (same signature length), we might be losing precision in the foveal representation with the new reduced vocabulary. To avoid this limitation, in the next section we reformulate our problem as follows: given a total signature length, and using saliency, we would like to optimally allocate the respective proportions for the foveal and peripheral channels.

5.2.2 Feature selection by Saliency-based Non-Uniform Sampling in a Variable-Resolution space

In this section we describe our approach towards the emulation of visual fields through the Non-Uniform Sampling of features at Variable Spatial Resolutions (NUS+VSR). As already mentioned, due to the varying morphology of neurons in the human retina, it simultaneously enables a high-resolution detailed perception in the visual field associated with the fovea, and a low-resolution one in the peripheral visual field [Wandell 95]. This leads to a non-uniform spatial resolution image analysis, commonly known as *foveation* [Chang 00].

Our approach combines space-variant sampling [Vig 12] and multi-resolution image foveation [Perry 02] to model differentiated visual fovea and peripheral pathways found in human retina [Wandell 95]. Fig. 5.2 illustrates the approach. Since we implement the non-uniform sampling as a pruning process that filters out many patches from an initial set, the first step of our approach is to define a dense grid of circular local patches of radius r (with a step size of 3px in our case). Let us note that this step just involves the definition of the grid (which is not costly at all) and that the computation of the descriptors (which requires an important computational burden) is just made after the pruning process is finished.

In order to simplify the model description, we split it into variable spatial resolution and non-uniform sampling.

5.2.2.1 VSR

In order to provide a multi-resolution analysis of an input image, we first discretize the resolution space by generating a multi-scale Gaussian image pyramid of L levels. Lower levels are meant to represent foveal vision whereas upper ones model peripheral vision. As shown in Fig. 5.2, we define the *resolution factor* ρ that stands for the ratio between the widths of two contiguous images in the pyramid $W_l = \frac{W_{l-1}}{\rho}$, where l denotes the level in the pyramid.

Then, using values in the saliency map, we can compute the saliency value of each local circular patch n as:

$$s_n = \underset{m \in \Omega_n}{max} (S(m)) \tag{5.3}$$

where $S(m)$ is the value of the saliency map normalized to the interval [0, 1] at location m and Ω_n stands for the pixels within the local patch n. We have found that max pooling here is more efficient for the target recognition task than mean pooling. Hence, depending on this value $s_n \in [0, 1]$, we assign each local patch to a particular level of the pyramid. In our case, this is done by a simple linear quantization ($Q(s)$ in Fig. 5.2) that uniformly splits

the resolution space into equally sized segments. Intuitively, based on the saliency value, we are modeling the foveal visual field as a high-resolution pathway that pays attention to small image details, whereas the peripheral vision path acquires information at a lower resolution, thus focusing on coarse visual patterns.

Let us note that we do not change the scale of the local regions (defined by the radius r); alternatively, we decrease the size of the image in each level so that the relative size of the local regions increases with the level l. Our approach therefore discretizes the resolution space, which differs from previous works towards foveated video displays [Chang 00, Perry 02], where continuous-resolution image representations (foveated images) were generated by interpolating previously computed discrete-resolution image representations. Although a continuous resolution space might seem appealing for our problem, its implementation leads to two problems which discourage its application: first, using interpolation between images at various resolutions and generating a foveated image based on the saliency map will lead to local regions containing pixels at various resolutions. This would produce the undesirable scenario in which local descriptors are computed over areas with non-uniform resolution. Indeed, it could be seen as a retinal visual cell working at variable resolution in its visual field, which does not coincide with our objective of modeling various visual cells, each of them working at a specific spatial resolution. In addition, computing image interpolations on-the-fly, ensuring that all pixels in a local patch correspond to the same resolution, would solve this issue at the expense of an important increase in the computational burden. As we will show in the experimental section, increasing the number of levels L (which, if we keep the resolution of the last level as a constant, would tend to a continuous resolution space if L is large enough) does not notably improve classification results.

5.2.2.2 Non Uniform Sampling (NUS)

In this section we introduce the pruning process that filters out non-relevant visual information in order to provide more compact image representations focusing on areas of high saliency.

We follow a similar approach to [Vig 12], in which a Weibull cumulative distribution was proposed to perform random sampling based on saliency values. In particular, defining a random variable S associated with visual saliency, the Weibull cumulative density function obeys:

$$F_S(s) = P(S \leq s) = 1 - e^{(-s/\lambda)^\kappa} \tag{5.4}$$

where κ is called the *shape* parameter and λ the *scale* parameter. Hence, for each n-th local region with a particular value s_n we randomly decide if it is pruned or not based on the value of $F_S(s_n)$.

Intuitively, the *shape* parameter κ controls the influence of the saliency value on the pruning process. Whereas low values of κ give less influence to the saliency value (both salient and non-salient areas have similar opportunities to survive the pruning process), high values will prune almost all the non-salient local regions in the final image representation. Further-

more, for a given κ value, the *scale* parameter λ controls the total amount of local regions being pruned.

We aim at improving classification results while avoiding any additional processing burden. Hence, in order to produce almost the same number of visual descriptors as in the uniform case, we have designed the following random sampling procedure. Let us consider a *desired number* of patches N, that corresponds with the number of processed patches in the baseline BoVW (no saliency). Since we are following a pruning process, our initial dense grid will produce a large enough set of N_0 points so that $N_0 >> N$. Then, we will set a value for the shape parameter κ in the Weibull distribution and, in order to keep the computational complexity constant, we will automatically calculate the corresponding λ value producing a final number of points $\hat{N} \sim N$.

For that end, let us consider $\hat{N}$ as a random variable and therefore compute its expected value as:

$$E[\hat{N}] = \sum_{n=1}^{N_0} 1 \cdot F_S(s_n) = N_0 - \sum_{n=1}^{N_0} e^{(-s_n/\lambda)^\kappa} \tag{5.5}$$

where we have considered that the probability of a patch n being included in the final image representation is the value of the Weibull cumulative density function $F_S(s_n)$ on the saliency of the patch s_n.

Unfortunately, from eq. (5.5) it is not possible to obtain an analytic optimal value of λ that makes $E[\hat{N}] = N$. However, since $s_n \geq 0$, $\lambda \geq 0$, and $\kappa \geq 0$, $g(s_n) = e^{(-s_n/\lambda)^\kappa}$ is a convex function over s_n. This allows us to apply Jensen's inequality to obtain an upper bound of eq. (5.5) as:

$$E[\hat{N}] \leq N_0 \left(1 - e^{-\frac{1}{N_0}\sum_{n=1}^{N_0}(s_n/\lambda)^\kappa}\right) \tag{5.6}$$

That is, we can obtain an upper bound of the number of points being processed, which allows us to successfully keep the computational complexity bounded for each value of κ. Working out λ in eq. (5.6) gives a final expression for λ:

$$\lambda = \left[\frac{E[s^\kappa]}{\ln\left(\frac{N_0}{N_0 - N}\right)}\right]^{1/\kappa} \tag{5.7}$$

where $E[s^\kappa] = \frac{1}{N_0}\sum_n s_n^\kappa$. The upper bound in (5.6) is tight when the values $(s_n/\lambda)^\kappa$ are very similar for every n. This means that we get better approximations $\hat{N} \sim N$ when κ is small than when it is very large (where we might get $\hat{N} < N$). As we will show in the experimental section, where we will cross-validate the value of κ, the influence of the approximation on the results is negligible and, in fact, better results are achieved for high values of κ.

5.2.3 SC

This section is devoted to the description of the third saliency-based stage of the pipeline: *Saliency-sensitive Coding of features*. As we already mentioned, this work is inspired by

previous proposals in which information belonging to foreground and background is encoded [de Carvalho Soares 12] independently, as well as by the principles of sparse coding [Yang 09] and locality coding [Wang 10]. However, we provide a self-organized approach that automatically learns the optimal vocabularies for each spatial resolution and then assigns each visual descriptor taking into account both its visual appearance and its associated saliency value.

To do so, we start by considering the Locality-constrained Linear Coding (LLC) approaches presented in [Wang 10] and [Wei 13]. In these works, some exponentially-increasing locality functions were used to provide sparse codes which represented a particular descriptor using a small subset of visual words from a vocabulary. In our case, while keeping the sparse requirement, we aim to provide a SC of features. The objective of SC is two-fold: first, we aim to generate particularized vocabularies for each spatial resolution so that each descriptor is coded as a linear combination of visual words acquired at close spatial resolutions; second, we aim to automatically set the optimal number of words assigned to each spatial resolution, so that more words are used to represent visually salient image locations and vice versa.

Our problem formulation is as follows: for a given set of N descriptors defined by the pair $\{\mathbf{x}_n, s_n\}$, where $\mathbf{x}_n \in \mathbb{R}^{Dx1}$ stands for the visual descriptor and s_n is the saliency value associated with the region (see eq. (5.3)), we define an over-complete visual vocabulary $\{B, \mathbf{p}\}$. The matrix $B \in \mathbb{R}^{DxK}$ contains the visual words of the vocabulary, whereas the vector $\mathbf{p} \in \mathbb{R}^{Kx1}$ defines the retinal path associated with each visual word. This path is a continuous variable in the range [0,1] that models a fuzzy membership to the foveal/peripheral pathways, in which 0 stands for a path completely associated with the peripheral vision (low spatial resolution) and 1 corresponds to the foveal path (high spatial resolution).

Hence, given a visual descriptor $\mathbf{x}_n$ computed at a particular spatial resolution (that depends on its associated saliency), we aim to represent it as a linear combination of a small set of visual words of the vocabulary, strengthening those of similar type (similar spatial resolution).

To that end, we formulate the following minimization problem:

$$\min_{\boldsymbol{\alpha},B,\mathbf{p}} \sum_{n=1}^{N} \left\{ \|\mathbf{x}_n - B\boldsymbol{\alpha}_n\|_2^2 + \lambda_l \|\mathbf{l}_n \odot \boldsymbol{\alpha}_n\|_2^2 + \lambda_t \|\mathbf{t}_n \odot \boldsymbol{\alpha}_n\|_2^2 \right\} \; s.t. \; \mathbf{1}^T \boldsymbol{\alpha}_n = 1 \tag{5.8}$$

where $\alpha_n \in \mathbb{R}^{Kx1}$ represents the vector of weights in the linear combination and is called the *code*, $\odot$ stands for the Hadamard product (element-wise) between two vectors, and $\mathbf{1}$ represents a vector of ones. The first element in eq. (5.8) corresponds to the coding error between the original and the reconstructed descriptor. The second element ensures locality by incorporating a *locality adaptor* $\mathbf{l}_n \in \mathbb{R}^{Kx1}$ to the problem. This locality adaptor, previously introduced in [Wang 10], stands for the visual distance l_{nk} between the descriptor and each word in the vocabulary. By using an exponentially-increasing adaptor of the form:

$$l_{nk} = \sqrt{\exp\left(\frac{\|\mathbf{x}_n - \mathbf{b}_k\|_2^2}{\sigma_l^2}\right)} \tag{5.9}$$

we are able to generate sparse codes α_n in which just a few α_{nk} associated with words that are close in the feature space get non-zero values. It is easy to notice that the lower the parameter σ_l^2, the sparser is the resulting code.

Finally, with the third term in eq. (5.8) we aim to code each descriptor using words in the vocabulary with similar spatial resolution. Therefore, we introduce a new *type* adaptor $\mathbf{t}_n \in \mathbb{R}^{Kx1}$ that compares the retinal paths of the descriptor and visual word as:

$$t_{nk} = \sqrt{\exp\left(\frac{\|s_n - p_k\|_2^2}{\sigma_t^2}\right)} \tag{5.10}$$

where, again, we have made use of an exponentially-increasing adaptor with its own parameter σ_t^2.

5.2.3.1 Approximate Inference

Since eq. (5.8) is independently convex in $\{\boldsymbol{\alpha}, B, \mathbf{p}\}$, we have followed a *coordinate descent - gradient descent* approach to find the optimal values. That is, by iteratively optimizing the functional with respect to each parameter, it is ensured that the algorithm will converge to a local minimum.

In particular, in order to provide a solution for the coding stage ($\boldsymbol{\alpha}$), we can rewrite eq. (5.8) as:

$$\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} \boldsymbol{\alpha}_n^T C_n \boldsymbol{\alpha}_n + \boldsymbol{\alpha}_n^T diag(\lambda_l \mathbf{l}_n^2 + \lambda_t \mathbf{t}_n^2)\boldsymbol{\alpha}_n + \eta(\mathbf{1}^T \boldsymbol{\alpha}_n - 1) \tag{5.11}$$

where we have defined a new matrix $C \in \mathbb{R}^{KxK}$, computed as $C = (\mathbf{x}_n \mathbf{1}^T - B)^T(\mathbf{x}_n \mathbf{1}^T - B)$. We have additionally converted the equality constraint over $\boldsymbol{\alpha}$ into a new term with a Lagrange multiplier η. The compressed notation $\mathbf{l}_n^2$ stands for a vector whose elements are the square of the elements in $\mathbf{l}_n$. Computing the derivative of this functional F w.r.t. $\boldsymbol{\alpha}_n$ gives:

$$\frac{\partial F}{\partial \boldsymbol{\alpha}_n} = 2C_n \boldsymbol{\alpha}_n + 2diag(\lambda_l \mathbf{l}_n^2 + \lambda_t \mathbf{t}_n^2)\boldsymbol{\alpha}_n + \eta \mathbf{1}^T \tag{5.12}$$

Setting this derivative to zero gives the following equation:

$$\begin{aligned} U\boldsymbol{\alpha}_n + \eta \mathbf{1} &= 0 \\ s.t.\ \mathbf{1}^T \boldsymbol{\alpha}_n &= 1 \end{aligned} \tag{5.13}$$

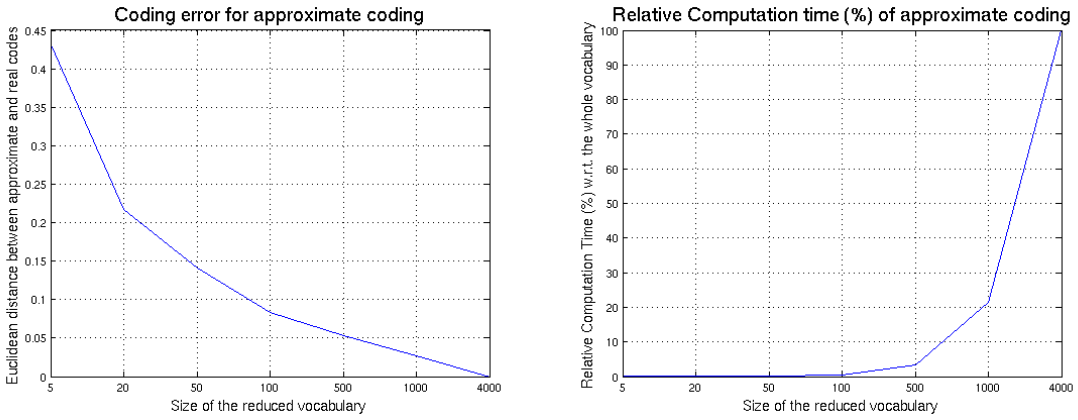
with $U = 2C + 2diag(\lambda_l \mathbf{l}_n^2 + \lambda_t \mathbf{t}_n^2)$.

Now, multiplying by $\mathbf{1}^T U^{-1}$ and applying the constraint ($\mathbf{1}^T \boldsymbol{\alpha}_n = 1$) we obtain:

$$\mathbf{1}^T \boldsymbol{\alpha}_n + \eta \mathbf{1}^T U^{-1} \mathbf{1} = 0 \tag{5.14}$$

Now we can get the value of the Lagrange multiplier η:

$$\eta = -\frac{\mathbf{1}}{\mathbf{1}^T U^{-1} \mathbf{1}} \tag{5.15}$$



(a) (b)

Figure 5.3 – a) Mean euclidean distance (mean Coding error) between the approximate codes and the real ones, and b) relative computation time needed to compute the codes for various sizes of the reduced vocabulary $\hat{K}$. In both cases, the complete vocabulary K is taken as reference.

And substitute in eq. (5.13) getting an expression for $\boldsymbol{\alpha}_n$:

$$\boldsymbol{\alpha}_n = \frac{U^{-1}\mathbf{1}}{\mathbf{1}^T U^{-1}\mathbf{1}} \tag{5.16}$$

Finally, defining $\tilde{\boldsymbol{\alpha}_n} = U^{-1}\mathbf{1}$, we get:

$$\boldsymbol{\alpha}_n = \frac{\tilde{\boldsymbol{\alpha}_n}}{\mathbf{1}^T \tilde{\boldsymbol{\alpha}_n}} \tag{5.17}$$

Unfortunately, exact inference becomes impractical when the size K of the vocabulary increases, as computing the inverse of the matrix $U \in \mathbb{R}^{KxK}$ is very computationally intensive. Hence, we have developed an approximate inference process as follows: (1) for each descriptor, we consider a reduced vocabulary of size $\hat{K} << K$, containing only those visual words k that minimize the partial functional $\lambda_l \|l_{nk}\|_2^2 + \lambda_t \|t_{nk}\|_2^2$; (2) then, we solve the simplified problem stated in eq. (5.11) for this reduced vocabulary.

In Fig. 5.3a and 5.3b we show the mean Coding error between the approximate codes and the real ones, and the computation time required by the coding process, respectively, for various sizes of the reduced vocabulary $\hat{K}$. From these results, we decided to take a reduced size $\hat{K} = 100$ as the final option as it provides a good trade-off between coding error and computation time.

For the p_k parameter, we need to solve the following unconstrained convex optimization problem:

$$\min_{\mathbf{p}} = \lambda_t \|\mathbf{t}_n \odot \boldsymbol{\alpha}_n\|_2^2 \tag{5.18}$$

It is easy to note that setting the derivative of (5.18) with respect to $\mathbf{p}$ equal to zero leads

to a nonlinear equation on $\mathbf{p}$. Hence, we can obtain an optimal value of $\mathbf{p}$ using a Newton-Raphson method that, in the iteration i updates $p_k^{(i+1)}$ as:

$$p_k^{(i+1)} = p_k^{(i)} + \frac{\sum_{n=1}^{N} \alpha_{nk}^2 \left(\mathbf{t}_{nk}^{(i)}\right)^2 \left(s_n - p_k^{(i)}\right)}{\sum_{n=1}^{N} \alpha_{nk}^2 \left(\mathbf{t}_{nk}^{(i)}\right)^2 \left(1 + \frac{2}{\sigma_t^2}\left(s_n - p_k^{(i)}\right)^2\right)} \tag{5.19}$$

Finally, since the term associated with the *type adaptor* does not depend on B, the dictionary can be updated by following the Newton method proposed in [Wei 13]. The interested reader is referred to that work for the complete derivation of the B update formulas.

It is worth noting that variables B and $\mathbf{p}$ are just learned in the dictionary building phase and remain fixed during the computation of the image signatures (when only the $\boldsymbol{\alpha}$ is computed). In addition, let us note that this method shows various open parameters, namely $\{\sigma_l, \lambda_l, \sigma_t, \lambda_t\}$. In the experimental section, we will show the influence of these terms and the optimal values for our problem.

5.3 Experiments and results

As discussed in the introduction, egocentric video content has been selected as our main benchmark for saliency-based active object recognition in video. Let us note that we do not aim to detect the presence of every object in the scene but of only those ones considered as 'active': e.g. those ones that are interacted by users, either manipulated or observed, and that become the main source of information to understand the scene. This particular problem fits well with the nature of saliency since, as we have already mentioned, it aims to drive the recognition process to the areas of interest in the image, therefore preventing from the detection of non-active objects that belong to the background of the scene.

5.3.1 Scenarios, Datasets and Evaluation Metrics

We consider two different scenarios in egocentric video: the first one is *constrained*; all the subjects perform actions in the same room and, therefore, interact with the same objects in the same context, e.g. a hospital scenario in which patients perform several activities. Here, the limited intra-class variation is only due to natural conditions: occlusions, lighting, etc. We have used two datasets to model this scenario: a) the publicly available *GTEA* dataset for Object Recognition [Fathi 11b], that contains cap-mounted videos showing 7 types of daily activities, each performed by 4 different subjects, and comprising 16 categories of manipulated objects; and b) the *Dem@acare* dataset, containing 27 videos, captured by a shoulder-mounted GoPro camera with real Alzheimer patients performing various instrumental daily activities in a controlled hospital environment. This dataset, generated under the Dem@care[1] research project, contains 18 categories of active objects.

The second scenario is *unconstrained*. The recordings are made at different locations and users interact with different instances of the same object categories. The intra-class variation

[1]Dem@acare Project: http://www.demcare.eu/

here is strong and the amount of training data is small (just a few instances of each object category). For this scenario, we have used the publicly available *ADL* dataset [Pirsiavash 12]. It contains videos captured by a chest-mounted GoPro camera on users performing various daily activities at their homes, showing objects from 44 categories. We have just considered objects labeled as 'active' in both training and testing. This dataset was used for two purposes: a) a reduced version was utilized to validate the free parameters of our proposal. Here a strong temporal sub-sampling was applied and the data was split into training and test sets with 1464/1251 frames respectively, ensuring that frames of the same video were not contained in both sets. And b) for the final evaluation, the 20 videos were divided into 5 sets of 4 videos each, so that a leave-one-out assessment was performed at this subset level.

As evaluation metrics, as presented in 3.4.2 and following the setup of the original authors, mAP was used for the Dem@acare and ADL datasets, and multiclass accuracy was applied in the GTEA dataset.

5.3.2 Validation of model parameters

Here we validate and show the influence on the system performance of every open parameter introduced in Sec. 5.2.

5.3.2.1 Variable Spatial Resolution and Non-Uniform Sampling

As we have four open parameters in this module, and although they are not independent, performing a joint fine cross-validation becomes impractical. Hence, after an initial very coarse parameter selection leading to an initial set of values, we have sequentially performed various 'one-at-a-time' optimizations with respect to:

1) *Radius r of the circular local regions*: in Fig. 5.4a we evaluate the influence of this parameter in two scenarios: with our VSR approach (blue), and (red) with a basic Dense Sampling approach corresponding to the baseline BoVW (red). Let us list the values of other parameters as: $\rho = 2$, $L = 4$ and NUS activated with $k = 5$ (Section 5.2.2.2). Whereas the optimal value for the basic dense grid was $r = 30$, for the VSR it was $r = 10$, as going up in the resolution pyramid increases the relative size of circular regions with respect to image dimensions. Furthermore, the better results achieved by the VSR scheme demonstrate its capacity for removing very fine details and thus focusing on coarser shapes at upper levels of the pyramid.

2) *The resolution factor ρ*: this factor relates the sizes of subsequent images in the resolution pyramid. In Fig. 5.4b we show the validation of this parameter, fixing the rest of the parameters ($r = 10$, $L = 4$ and NUS activated with $k = 5$). Good results are obtained in the range $\rho \in [1.5, 2]$ so that we have set $\rho = 1.5$ as the final one.

3) *The number of levels in the pyramid L*: this parameter controls the degree of discretization of the resolution space. As discussed in Sec. 5.2.2, in order to keep this complexity as bounded as possible, we work with a discretized resolution. To isolate the influence of L from ρ, we have fixed W_{L-1}, the smallest size of the image in the top level of the pyramid, with independence of the number of levels L; and then, for each evaluated L, we have

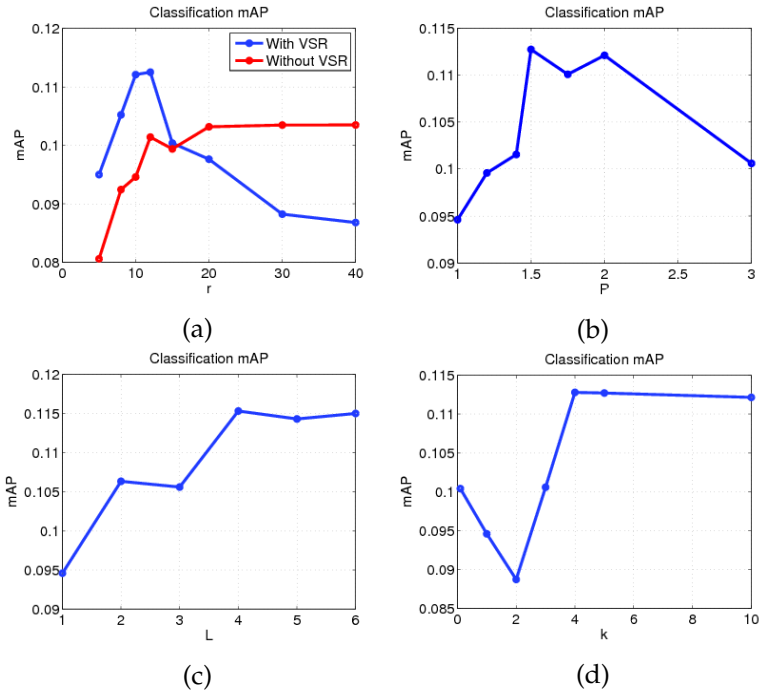


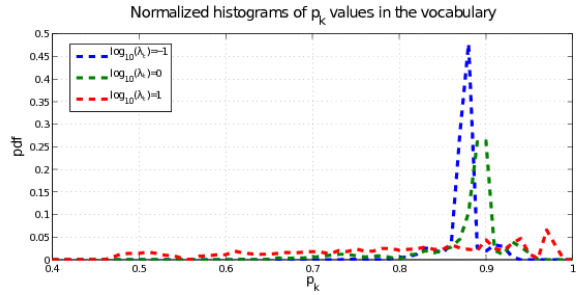
Figure 5.4 – Validation of the VSR+NUS parameters in the ADL unconstrained dataset: (a) radius r of the circular regions, (b) resolution factor ρ, (c) number of levels in the pyramid L, (d) shape parameter k in the Weibull distribution of the NUS.

accordingly set the resolution factor ρ that produces this W_{L-1} in the top level of the pyramid. The results provided in Fig. 5.4c demonstrate that the performance grows until $L = 4$, from which it stabilizes and more continuous representations of the resolution space do not improve the performance.

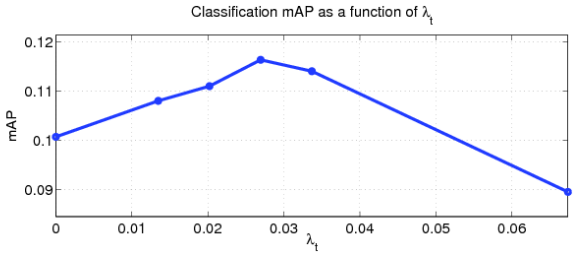
4) *The shape parameter k of NUS*: for each value of k in the NUS module, we have accordingly computed the scale value λ that produces the desired final number of points $\hat{N} \sim N$ (see Sec. 5.2.2). The results of this study are presented in Fig. 5.4d, and show that larger values of k are preferred in the sampling process (we get an optimal value of $k = 5$). For these values, the sampling process is highly unbalanced so that many more points are detected in high-saliency areas than in low ones.

5.3.2.2 SC

Since our Saliency-Sensitive Coding is built over LLC [Wang 10, Wei 13], we have firstly set the values for the locality adaptor. As it is not the scope of this chapter, we do not include figures about their influence but simply note that the optimal values were $\lambda_l = 0.10$, $\sigma_l^2 = 0.25$. On the contrary, we are interested in the study of the influence of the saliency-based *type adaptor* in the coding process. Two are the parameters of this adaptor (Sec. 5.2.3): λ_t, which controls the weight of the saliency over the coding process, and σ_t, that handles the degree of nonlinearity of the coding process with the saliency.



(a)



(b)

Figure 5.5 – (a) Normalized histograms (pdf) of the p_k in the visual vocabulary for various values of λ_t. (b) Validation of λ_t parameter of the SC in ADL dataset.

The influence of the saliency type adaptor is illustrated in Fig. 5.5a. For several values of the parameter λ_t, we show the normalized histograms of the p_k values in the visual dictionary. As can be seen, when λ_t is small, the visual path of the visual words is ignored in the coding process so that the values of p_k tend to the sample average of the descriptors' saliency s_i (histograms with a sharp peak around this sample average saliency value of descriptors in the dataset). In contrast, if λ_t is high, each word in the vocabulary is associated with a particular visual retinal path and, consequently, with a smaller range of saliency values, thus giving place to a more uniform histogram. Note that we present here a process in which the number of visual words devoted to foveal and peripheral paths is automatically learned from data. Furthermore, the assignment of a descriptor to a particular path is performed softly, and depends on the distance between the saliency of the descriptor s_n and the type value p_k of each word in the vocabulary. In Fig. 5.5b, we show additional results of our validation of the λ_t parameter for a heuristically computed optimal σ_t value of $\sigma_t^2 = 0.1$.

5.3.3 Comparing Saliency Approaches

To assess the influence of each saliency-based stage in the active object recognition problem, we have compared several versions of our approach, namely:
a) *Reference methods*:

1. *BoW*: Baseline BoVW with a vocabulary size of 4000 visual words.

2. *BoW + GT Masks*: This approach utilizes (human-annotated) Ground Truth bounding boxes of the active objects, and filters out the descriptors associated with local regions

Table 5.1 – A comparison of various configurations of Saliency-based Object Recognition for the whole (44) and reduced (10) sets of categories in the ADL dataset: mAP and p-value of a paired t-test taking NUS +VSR + SC as reference.

Algorithm/ mAP (p-value)	ADL (44 cat)	ADL (10 cat)
BoW	13.6 (0.08)	32.8 (0.02)
BoW + GT Masks	16.8 (0.77)	50.4 (0.47)
NUS [Vig 12]	13.9 (0.03)	35.1 (0.00)
NUS+VSR	15.1 (0.14)	38.3 (0.00)
SP-F	14.5 (0.06)	39.4 (0.04)
SP-FP (2000) [de Carvalho Soares 12]	13.8 (0.11)	35.6 (0.11)
SP-FP (4000) [de Carvalho Soares 12]	14.7 (0.25)	38.2 (0.19)
NUS+VSR+SP-F	14.2 (0.01)	37.5 (0.01)
NUS+VSR+SC	**16.2**	**44.0**

located outside the objects of interest.

b) *Visual Fields with VSR and NUS*:

1. *NUS*: BoVW with our NUS module described in sec. 5.2.2.2, that extends the work in [Vig 12].
2. *NUS+VSR*: We add the VSR module (sec. 5.2.2.1) to the previous approach.

c) *Saliency Pooling Methods*:

1. *SP-F*: BoW + Saliency Pooling considering only the foveal weights as stated in eq. (5.1).
2. *SP-FP*: BoW + Saliency Pooling considering both the contributions to the foveal and peripheral vision, as stated in eq. (5.2) and [de Carvalho Soares 12]. We have tested two vocabulary sizes: 2000 words (keeping the length of the image signatures constant), or 4000 (doubling the length of the image signatures).

d) *Combined methods*:

1. *NUS+VSR+SP-F*: We add Foveal Saliency Pooling to the NUS+VSR version of our approach to study the combination of both.
2. *NUS+VSR+SC*: We substitute the Saliency pooling by our Saliency-sensitive Coding with a vocabulary size of 4000 words.

Results for every method are shown in Table 5.1. Regarding the methods implementing the VSR+NUS visual fields, we appreciate the positive influence of both elements in the results: although the NUS already enhances the basic BoVW, the VSR approach yet provides notable improvements to system performance. This is a consequence of the spatial resolution

adaptation to the foveal and peripheral vision, and raises the need for independent and different scale processing paths for the objects of interest (active objects) and their surrounding context.

In parallel, SP also helps to enhance system performance, even if we merely model the foveal vision (SP-F). The approach in [de Carvalho Soares 12], which concurrently models foveal and peripheral vision obtains varying results depending on the scenario: if we aim to keep the computational complexity constant and then reduce the vocabulary to half of the size (2000 words), we get a dramatic loss in performance. This result might be expected as we are adding a new path with peripheral vision but at the expense of decreasing the precision of foveal vision. Although we consider that peripheral vision provides useful information, giving the same weight to both pathways has turned to be inappropriate. Furthermore, the fact that keeping the vocabulary size constant (image signatures of double length) improves the performance demonstrates that modeling context is also useful due to the general correlation between objects and locations.

Finally, the combination of various saliency-aware approaches reveals disparate results. Whereas we have observed that the combination of variable resolution visual fields and saliency pooling has not improved performances, saliency coding successfully combines with other saliency-based modules in the processing pipeline. From our point of view, the rationale behind this is that the automatic approach of saliency-sensitive coding correctly handles the relative importance of foveal and peripheral visual pathways, even in the presence of previous blocks in the processing pipeline (like the visual fields described in Sec 5.2.2). This is something that does not occur with Saliency Pooling which, although by itself provides good performance, when combined with other modules weighs in excess the foveal with respect to the peripheral path and therefore cancels the influence of the context in the recognition process.

Due to these results, in the following, we will assess the best performing approach *NUS+VSR+SC* in other scenarios and in comparison with various methods that have reported state-of-the-art results in the considered datasets.

5.3.4 Comparison with the State-of-the-Art

In Table 5.2 we include a comparison between our approach (denoted as 'Ours'), the reference methods, and some techniques that reported State-of-the-art results in egocentric vision and other object detection datasets: a) the discriminatively-trained Deformable Part Model (DPM) [Felzenszwalb 10], a sliding window technique that has reported the state-of-the-art results for the ADL dataset [Pirsiavash 12]; and b) the object recognition approach designed and reported by the authors of the GTEA dataset [Fathi 11b]. In addition, for the ADL dataset, we have also evaluated two methods that combine well-known object recognition approaches and saliency, namely: c) DPM over bounding boxes proposed in [Alexe 12] (DPM + obj.), which randomly samples bounding boxes and selects the most appropriate based on a measure of their objectness (likelihood to contain an object); and d) BoVW applied over candidate bounding boxes proposed by the method described in [Uijlings 13] (BoW + SS), which applies a SS to generate potential candidate locations for objects. In both cases, we

Table 5.2 – A comparison between our method and some state-of-the-art approaches for various datasets. The p-value of a paired t-test taking 'Ours' as reference is included when available.

	Constrained		Unconstrained	
Dataset	GTEA	Dem@	ADL(44)	ADL(10)
Algorithm/Metrics	Acc	mAP	mAP	mAP
BoW	35.0	45.3 (0.17)	13.6 (0.08)	32.8 (0.02)
BoW + GT Masks	-	54.8 (0.53)	16.8 (0.77)	50.4 (0.47)
DPM [Felzenszwalb 10]	-	34.9 (0.01)	15.3 (0.61)	42.4 (0.66)
DPM [Felzenszwalb 10] + obj. [Alexe 12]	-	-	13.1 (0.06)	35.9 (0.05)
BoW + SS [Uijlings 13]	-	-	13.4 (0.08)	36.9 (0.19)
Fathi et al. [Fathi 11b]	35.0	-	-	-
Ours	**45.4**	**50.9**	**16.2**	**44.0**

have followed the same setup described in the original papers [Alexe 12, Uijlings 13] to develop the object detector. However, in order to establish a fair comparison with our method, in BoW + SS we have used the same features (dense SURF features) of our proposal.

As we can see from the results, our method consistently outperforms any other automatic approach in every dataset, as well as achieves close performance to the hypothetic case in which ground truth masks/bounding boxes are available (GTEA lacks Ground Truth in all videos so we cannot provide BoW + GT results for this dataset). In particular, in the GTEA dataset we achieve absolute improvements of 10% compared to the best reported approach for this dataset [Fathi 11b]. The performance of DPM is significantly worse than ours under the constrained scenario (Dem@care dataset), whereas it gets closer results under the unconstrained one (ADL dataset). The rationale behind this is that the design of this method is intended to provide good generalizations of the objects' appearance. This property, although desirable under unconstrained scenarios, leads to a loss in performance for the detection of particular object instances (constrained scenario). The very high p-value between DPM and our approach in the unconstrained scenario means that the improvement of our method is not consistent through all the categories, and that DPM is the best choice for some of them (see Fig. 5.6).

Furthermore, the two state-of-the-art approaches using saliency show lower performance as they strongly restrict the set of locations and scales to be evaluated by the detectors. In particular, the restricted set of candidate windows in DPM+obj causes non-detections with respect to the full DPM approach, whereas for the BoW+SS we have found that, although learning object appearance from accurate ground truth bounding boxes may provide additional information such as accurate object localization, it is very sensitive to the quality of the automatically proposed boxes in test images.

In Fig. 5.6 we also include per-category results in the ADL dataset. Together with the

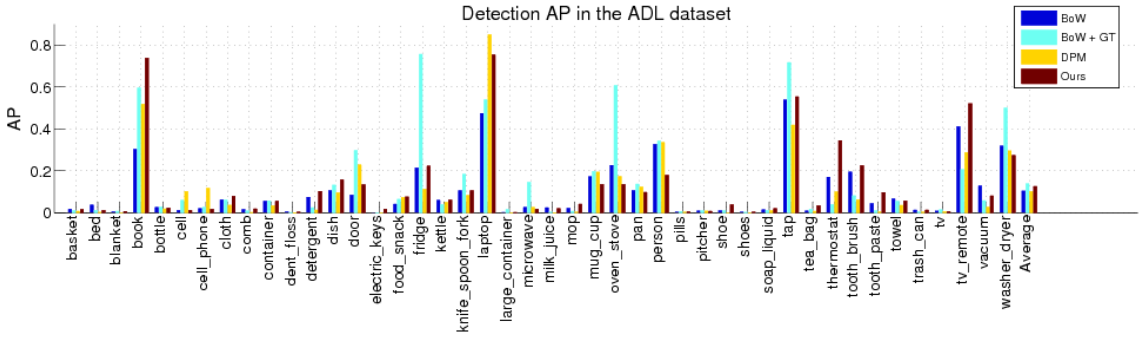


Figure 5.6 – Detailed per-category results of various approaches in ADL dataset

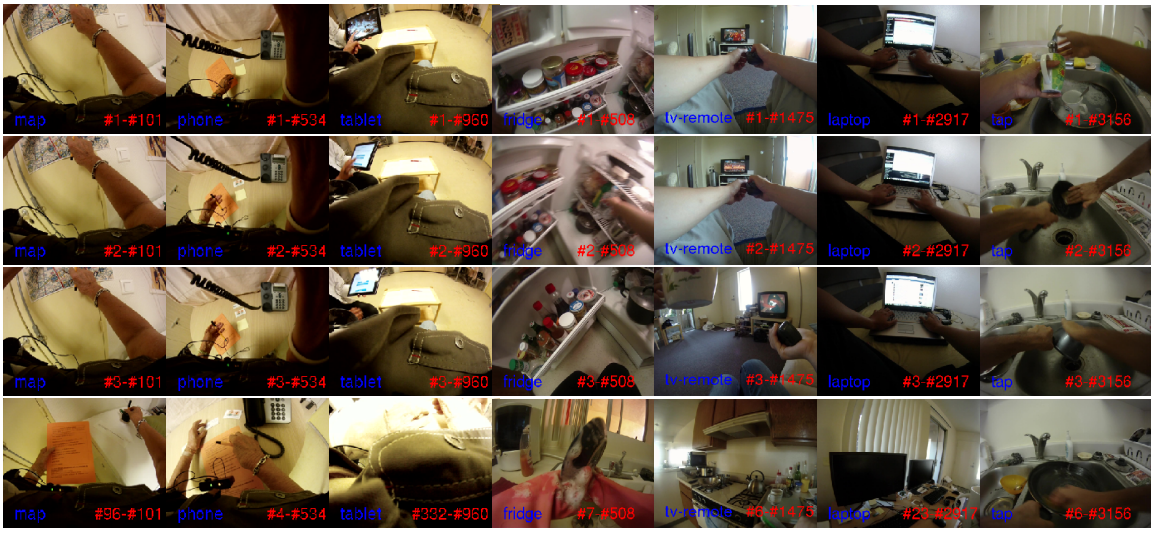


Figure 5.7 – Some visual examples of the ranking provided by our system. Each column represents an object category (columns 1-3 Dem@care dataset, columns 4-7 ADL dataset). For each sample we show the top three ranked results and the first non-relevant ranked image, including the #Ranking Position - #Number of relevant images.

visual examples in Fig. 5.7, they yield additional conclusions. In general, the results vary significantly from one class to another. The poor results in some categories can be explained as follows: although the number of image samples of a class may be high enough (hundreds of thousands), they correspond to a small set of different object instances (no more than 10-15 different instances per category). Hence, if a category shows high intra-class variation (e.g. bed, blanket, container or cloth), it is not possible to obtain good generalizations with such limited training sets. Although using external databases might seem appealing, the work in [Pirsiavash 12] showed how the application of detectors trained in ImageNet [Deng 09] yielded poor results for this particular dataset.

For categories in which the BoW + GT achieves notably better results than the automatic approaches, we have observed that some of the errors are found in images in which the object is present but not considered as active. This particularly holds for static objects which,

Table 5.3 – Comparison of S.T. and M.T. execution times.

Case	DPM [Felzenszwalb 10]	Proposal w/o SC	Proposal w SC
S.T.	60.4s	6.7s	15.1s
M.T.	10.9s	2.6s	6.0s

although can be manipulated by humans, are rarely moved (e.g. phone and tap belong to the context in the last row in Fig. 5.7). In those cases only the GT bounding boxes guide the recognition process exactly to the active object, which is hardly achieved by any automatic saliency method. We have studied the effect of this kind of error by removing from the evaluation those frames where an object is present but not active and concluded that it equally affects all compared algorithms. Conversely, if the object of interest is very small (e.g. tv_remote in Fig. 5.7, thermostat, tooth_brush, etc.), the GT boxes do not contain enough discriminative information whereas our automatic saliency-based approach considers both the salient area and its context, therefore enhancing the detection of the object.

5.3.5 A study of the computation time

In Table 5.3, we show a comparison between the average execution times of our proposal and the DPM to run one category object-detector in a test frame. We include results using a single threading (S.T.) and multi-threading (M.T.) in a 2.10GHz computer with 4 cores and hyper-threading. For our proposal, the execution time comprises the whole processing pipeline shown in Fig. 5.1. It is worth noting that some of the computations for the spatial saliency map are implemented in GPU so they cannot be translated to S.T. case (spatial saliency takes about 0.05 sec per frame in the GPU). The rest of the calculations are made with the CPU under the aforementioned circumstances. For the DPM, we run the implementation in [Felzenszwalb 10], made in Matlab with optimized c routines for all the steps in the process that require most of the execution time. Our approach shows much lower computational times in comparison with DPM. The rationale behind this is the fact that using the saliency maps, we avoid the heavy scanning process of a sliding window approach such as the DPM. Furthermore, Saliency Coding becomes an important source of overhead in the execution time but, as we have shown in the experimental section, it also achieves a notable enhancement of the performance.

In our experiments, as we kept constant the number of features ($\hat{N} \sim N$ in Sec. 5.2.2), the computational complexity of BoW is similar to our method without SC (in except for the saliency map calculation). However, if the goal is to decrease the complexity, we could aim to obtain similar performances as BoW, and consequently strongly reduce the number of feature points and the computational complexity.

5.4 Conclusions

The application of saliency to computer vision has been traditionally restricted to a pre-processing stage that filters out non-relevant areas of an image. In this chapter, instead, we have proposed a perceptual model that incorporates visual attention to the challenging task of active object recognition in video and images. To do so, we have modeled independent foveal and peripheral pathways found in human retina, with particular properties in terms of spatial location, resolution, or sampling. In particular, we have introduced saliency into three particular processing modules of the well-known BoVW paradigm: a) Visual Fields with Variable-Resolution and Non Uniform Sampling, b) Saliency-sensitive Coding of features, and c) Saliency-based Pooling.

In order to assess the performance of our approach, we have selected the egocentric video as our main experimental benchmark. After discussing the influence of each module and its parameters, we have shown how our biologically-inspired saliency-based model helps to enhance current system performance. It not only achieves notable improvements with respect to the baseline BoW, but also provides state-of-the-art results in all the considered egocentric datasets at very competitive computational times. Furthermore, it avoids human efforts devoted to bounding-box level database annotation as in both training and test sets the saliency maps are automatically computed. A limitation of our method is revealed in scenarios where all present objects (active and non-active) have to be identified. In this case our saliency maps remove important visual information and restrict the performance of our approach. Another limitation are the performances of the BoVW model itself. Indeed, even if we manage to achieve state-of-the-art performances at a great computation rate, we can see that we clearly get close from the highest possible results obtained by the BoVW model, that is to say the case when ground truth masks locations are provided. In the future, we aim to continue exploring novel ways to introduce perceptual modeling into other object recognition models. It would be interesting indeed to see how visual attention modeling could improve the new best in class object recognition paradigms, for example using deep learning architectures.

The next chapter of this manuscript introduces our last contribution in which we propose to integrate our saliency-based active object recognition framework in an activity recognition model and study the performances.

Chapter 6

Application of object recognition to IADL recognition

6.1 Introduction

The task of recognizing human activities in videos has become a fundamental challenge among the computer vision community[Gaidon 09]. In order to face the limited field of view and the difficulty of accessing all relevant information from fixed cameras, an alternative has been found in egocentric videos, recorded by cameras worn by subjects. Indeed, in addition to dealing with the previously listed drawbacks, wearable cameras represent a cheap and effective way to record users activity for scenarios such as telemedicine or life-logging.

In this chapter we focus on the problem of recognizing IADLs for the assessment of the ability of patients suffering from Alzheimer disease and age-related dementia. Indeed, an objective assessment of a patient's capability to perform IADLs is a part of clinical protocol of dementia diagnostics and evaluation of efficacy of therapetical treatment [Amieva 08]. Traditionnal ways of assessment with the help of questionnaires do not bring satisfaction as two kinds of errors have been observed, which do not allow a practitionner to fully trust the responses. The error of the first kind is that one commited by the patients. At the early stage of dementia, they cannot admit that they become less performant in their everyday activities and diminish their difficulties. The error of the second kind is commited by the caregives. They are permanently stressed whatching their relatives' mental capacities to deteriorate. Hence they over estimate the difficulties of patients with dementia[Helmer 06]. This is why the egocentric video has been first used for the recording of IADLs on patients with dementia in [Mégret 10]. Later on, first results of recognition of IADLs in such recorded video were reported in [Karaman 14]. Nevertheless, the recognition problem being very complex, efficient ways of solving it still remain an open research issue.

There has been a fair amount of work on recognizing everyday at home activities by analyzing egocentric videos, many of them based on the fact that manipulated objects represent a significant part of the actions. However, most of the studies were conducted under a constrained scenario, in which all the subjects wearing the cameras perform actions in the same room and, therefore, interact with the same objects: e.g. a hospital scenario in which the medical staff asks patients to perform several activities. Typical constrained scenarios allow to make assumptions on the objects or even to use instance-level visual recognition: the authors in [Fathi 11a] present a model for learning objects and actions with very little supervision, whereas in [Sundaram 09] a dynamic Bayesian network that infer activities from location, objects and interactions is proposed. The problem still open under such a scenario becomes even more complex, if an ecological observation is performed, i.e. at person's home. The individual environment varies, the objects of the same usage, e.g. a tea-pot or a coffee machine, can be of totally different appearance. We call this scenario "unconstrained". In this case the recognition of activities in a wearable camera video has to be funded on the features of higher abstraction level than simple image and video descriptors computed from pixels.

It is only recently that the more challenging unconstrained scenario has been examined regarding activity recognition, such as in the work of [Kitani 11], where the authors recognize ego-actions in outdoor environments using a stacked Dirichlet Process Mixture model. Pirsivash and Ramanan [Pirsiavash 12] propose to train classifiers for activities based on the output of the well-known deformable part model [Felzenszwalb 10] using temporal pyramids. They demonstrate that performances are dramatically increased if one has knowledge of the object being interacted with. The approach making use of these "active" areas for ADL recognition has also been studied by Fathi and al. in [Fathi 12, González Díaz 13] under a constrained scenario, where the authors enhanced their performances by defining visual saliency maps.

In the context presented in this thesis that is to say the medical research on Alzheimer disease, the unconstrained scenario means an epidemiological study of performances of patients in an ecological situation at their homes as it was done in [Karaman 10]. Hence in this chapter we model an activity as a combination of a meaningful object the person interacts with and the environment. The rationale here is quite straightforward. Indeed a reasonable assumption can be made that e.g. if a person is manipulating a tea pot in front of a kitchen table, than the activity consists in "making tea". If a TV set is observed in the camera view field and the person is in living room, then the activity would be "watching TV". Therefore, efficient recognition approaches have to be proposed for object recognition and localization of a person in its environment, and, more than that an efficient combination of results of these two detectors have to be designed in the activity recognition framework.

This contribution shows that analyzing the dynamics of a sequence of active objects + context by means of temporal pyramids [Pirsiavash 12] becomes a suitable paradigm for activity recognition in egocentric videos. However, in this optic we claim that context can be better described by the output of place recognition module rather by the outputs of many non-active object detectors as proposed in [Pirsiavash 12]. We provide experimental evaluation on a publicly available dataset of activities in egocentric videos.

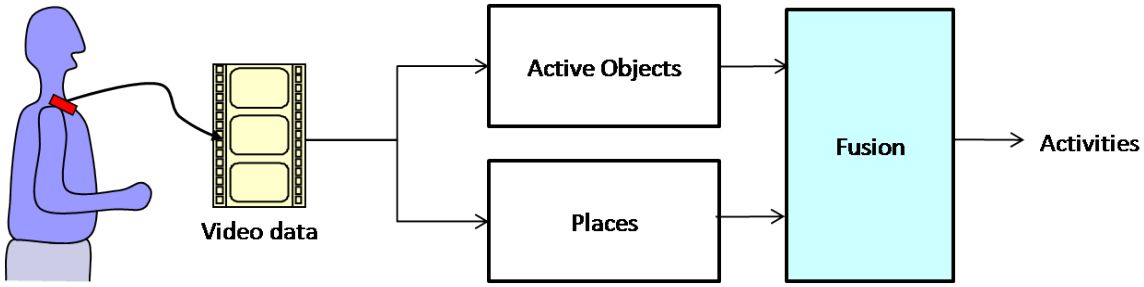


Figure 6.1 – Processing pipeline for the activity recognition

The remainder of the chapter is organized as follows: in Section 6.2 we describe the involved modules in our activity recognition approach. Section 6.3 assesses our model and compares it to the current state-of-the-art performances and section 6.4 draws our main conclusions and introduces our further research.

6.2 The approach

We aim to recognize IADLs by analyzing human-object interactions and as well as the contextual information surrounding them. Hence let us firstly introduce the notion of an *active object*. An active object is an object which the subject /patient wearing camera interacts with. Here the interaction is understood as manipulation or observation. We claim that the analysis of this kind of objects becomes the main source of information for the activity recognition, and that the explicit recognition of 'non-active' objects as in [Karaman 10] is not longer needed. We suppose that they can be efficiently encoded in a global descriptor of the scene/-context. The activity model is therefore understand as the interaction with active objects in a specific environment (context). In this particular work, we have considered that context can be successfully represented by identifying the place in which the user is performing the activity.

We propose a hierarchical approach with two connected processing layers (see figure 6.1). The first layer contains a set of *Active Object detectors* (Sec. 6.2.1) and a *Place Recognition* system (Sec. 6.2.2). Hence it allows for identification of the elements of our activity model. The second one addresses the activity recognition task on the basis of identified elements (Sec. 6.2.3).

6.2.1 Object Recognition

As already mentioned, we aim to recognize activities under an unconstrained scenario in which each video is recorded at a different place. This is therefore a more difficult task than the recognition of specific objects instances. It remains an open problem for the computer vision community.

For the rest of this study, the object recognition framework we employed is the one presented in section 5.2.1 using the BoVW combined with a saliency pooling framework. In

summary, we start by computing spatio-temporal bottom-up saliency maps for each frame based on the method from section 4.1.1 with central bias. Then once each image is represented by its weighted histogram of visual words, according to the method presented in 5.2.1, we use an SVM classifier [Cortes 95] with a nonlinear χ^2 kernel, which has shown good performance in visual recognition tasks working with normalized histograms as those ones used in the BoW paradigm [Sivic 03]. Using the Platt aproximation [Platt 99], we finally produce posterior probabilistic estimates O_k^t for the occurrence of the object of class k in the frame t.

6.2.2 Place Recognition

In this section we detail the place recognition module. Place recognition plays a role of *context recognition* in our overall approach for IADLs modeling and recognition.

The general framework was developed by the research team at the IMS laboratory [Wannous 12] and can be decomposed into three steps. First of all, for each image, a global image descriptor is extracted. They choose the Composed Receptive Field Histograms (CRFH) [Pronobis 10] since it was proven to perform well for indoor localization estimation [Dovgalecs 13]. Then a non-linear dimensionality reduction method is employed. In this case, they use a Kernel Principal Component Analysis (KPCA) [Schölkopf 98]. The purpose of this step is twofold: it reduces the size of the image descriptor which alleviates the computational burden of the rest of the framework, and it provides descriptors on which linear operations can be performed. Finally, based on these features, a linear SVM [Cortes 95] is applied to perform the place recognition, and the result is regularized using temporal accumulation [Dovgalecs 13].

For the application considered in this work, each video is taken in a different environment. Consequently, the module has to learn generic concepts instead of specific ones as it is usually the case [Dovgalecs 13]. In this context, we need to define concepts both relevant for action recognition and as constrained as possible to obtain better performances. Indeed, for example the concept 'sink area' has probably less variability and may be more meaningfull for action recognition than the concept 'kitchen'. This will be discussed in detail in Sec. 6.3.3.

Again, following the Platt aproximation [Platt 99], the output of this module is then a vector P_j^t with the probability of a frame t representing the place j.

6.2.3 Activity Recognition

Our activity recognition module uses the temporal pyramid of features presented in [Pirsiavash 12], which allows to exploit the dynamics of user's behaviour in egocentric videos. However, rather than combining features for active/non-active objects, we represent activities as sequences of active objects and places (context). For instance, cooking may involve user's interaction with various utensils whereas cleaning the house might require a user to move around various places of the house.

In particular, for each frame t being analyzed, we consider a temporal neighborhood Ω_t corresponding to the interval $[t - \Delta/2, t + \Delta/2]$. This interval is then iteratively partitioned

into two subsegments following a pyramid approach, so that at each level $l = 0 \dots L-1$ the pyramid contains 2^l subsegments. Hence, the final feature of a pyramid with L levels is defined as:

$$F_t = \left[F_t^{0,1} \dots F_t^{l,1} \dots F_t^{l,2^l} \dots F_t^{L-1,2^{L-1}} \right] \tag{6.1}$$

where $F_t^{l,m}$ represents the feature associated to the subsegment m in the level l of the pyramid and is computed as:

$$F_t^{l,m} = \frac{2^l}{\Delta} \sum_{s \in \Omega_{tm}^l} f_s \tag{6.2}$$

where Ω_{tm}^l represents the m temporal neighborhood of the frame t in the level l of the pyramid and f_s is the feature computed at frame s in the video. In the experimental section, we will assess the performance of our approach using the ouputs of K object detectors $[O_1^s \dots O_K^s]$, the outputs of J place detectors $[P_1^s \dots P_J^s]$, or the concatenation of both, as features f_s (in eq. 6.2).

In this work, we have used a sliding window method with a fixed window of size Δ, parameter that is later studied in the Sec 6.3, and a pyramid with $L = 2$. Finally, the temporal feature pyramid has been used as input for a linear multiclass SVM in charge of deciding the most likely action for each frame.

The complexity of the classifier system, being layered, precludes the easy interpretation of the results as probabilistic elements, as they are defined on an arbitrary axis that is suitable for deciding of a best class, but not to associate a probabilistic interpretation to it. Since automatic activity recognition from wearable camera is a difficult problem, it is very important to be able to assign confidence measures to these predictions, in order to monitor their validity and uncertainty for higher level inference. This problem corresponds to a calibration problem [Gebel 09]. Even though the automatic detection of all possible events is not possible in all cases, computing confidences can mitigate this, by trusting the prediction only when the system is confident.

For a two-class classifier, each observation x_k (in our case the input fatures belonging to a multi-dimensional space) is associated a predicted binary label y_n in $\{0, 1\}$. In practice the prediction is based on the thresholding of the classifier score s_n, which is produced by the decision function as $s_n = f(x_n)$. The calibration problem consists in finding a transformation $p_n = g(s_n)$ of these scores into a value in the interval $[0, 1]$ such that the result can be interpreted as the probability $p_k = P(y_k = 1|x_k)$ that of a true positive conditioned on the observed sample. The calibrated values have then reasonable properties to be used in a fusion approach with other sources of information.

In our work, we used the Platt approach [Platt 99], generalized to the one-to-one multiclass classification [Wu 04] and detailled in [Chang 11]. Each test sample is therefore associated with probabilistic confidence value $p_{kc} = P(L_k = c|x_k)$ that it belongs to class c, such that it is normalized by $\sum_c p_{kc}$.

The experimental part will evaluate both the raw recognition performance, using the classification strategy that assigns a sample to the class with higher probability, as well as the reliability of the estimated confidence value.

Figure 6.2 – Overview of the 18 activities annotated in the ADL dataset

6.3 Experimental Section

6.3.1 Experimental Set-up

We have assessed our model in the ADL dataset, proposed by the authors of [Pirsiavash 12], that contains videos captured by a chest-mounted GoPro camera on 20 users performing various daily activities at their homes. This dataset was already annotated for 44 object-categories and 18 activities of interest (see figure 6.2) and we have additionally labeled 5 rooms and 7 places of interest.

This dataset is very challenging since both the environment and the object instances are completely different for each user, thus leading to an unconstrained scenario. Hence, and due to the hierarchical nature of the activity recognition process, we have trained every module following a leave-k-out procedure (k=4 in our approach). This approach allows us to provide real testing results in object and place recognition for every user, so that the whole set can be later used for activity recognition. Furthermore, for activity recognition, the first 6 users have been taken to cross-validate the parameters of a linear SVM [Cortes 95], whereas the remainder ones (7-20) have been used to train and test the models following a leave-1-out approach. The library libSVM [Chang 11] was used for the classification.

6.3.2 Object recognition results

Figure 6.3 shows the per-category and average results achieved by our active object detection approach in terms of AP. We have used this quality measure rather than accuracy due to the nature of the dataset, which is highly unbalanced for every category. The mAP of our approach is 0.11 but, as can be noticed from the figure, the performance notably differs from one class to another. Main errors in classification are due to various reasons: a) a high degree of intra-class variation between instances of objects found at different homes, what leads to poor recognition rates (e.g. bed clothes or shoes show large variations in their appearance), b) some objects are too small to be correctly detected (dent floss, pills,

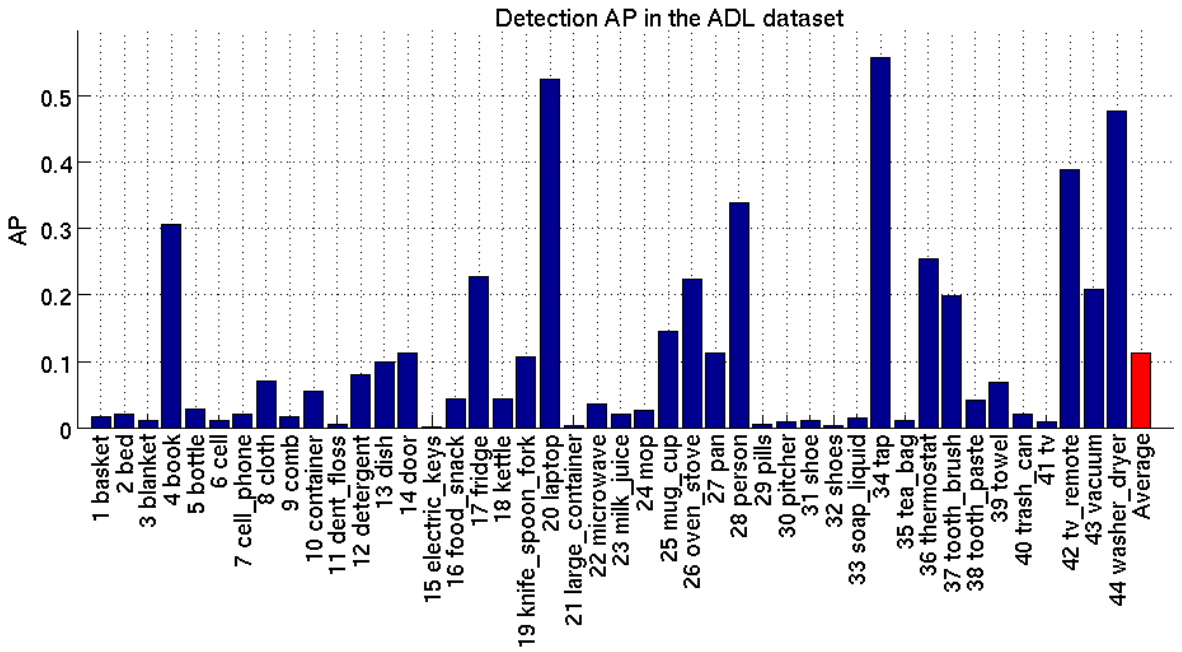


Figure 6.3 – Results in object detection

etc.), and c) for some objects that theoretically show a lower degree of intra-class variation (TV, microwave), performance is lower than expected since it is very hard for a detector to distinguish when they can be considered as 'active' in the scene (e.g. a user just faces a 'tv remote' or a 'laptop' when using them, whereas the TV or the microwave are more likely to appear in the field of view even when they are not ' active' for the user).

6.3.3 Place Recognition results

In this section, we report the results obtained on the ADL dataset for the place recognition module. We use a χ^2 kernel and retain 500 dimensions for the KPCA. We compared two different types of annotation of the environment: a room based annotation compound of 5 classes (bathroom, bedroom, kitchen, living room, outside) and a place based annotation compound of 7 classes (in front of the bathroom sink, in front of the washing machine, in front of the kitchen sink, in front of the television, in front of the stove, in front of the fridge and outside).

We have obtained average accuracies of 58.6% and 68.4%, for the room and place recognition, respectively. We will consider both features as contextual information for the recognition of activities.

6.3.4 IADLs Recognition results

In this section we show our results in IADLs recognition in egocentric videos. As already mentioned, our system identifies the activity at every frame of the video using a sliding

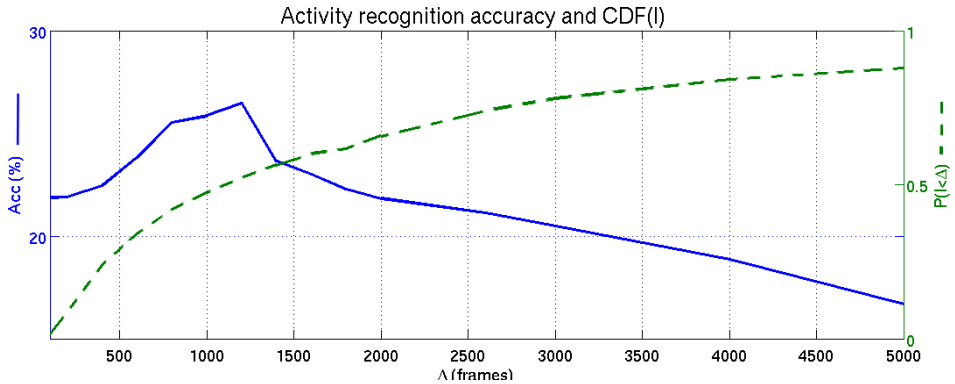


Figure 6.4 – Activity Recognition Accuracy with respect to the window size Δ (blue) and Cumulative distribution of activity lengths (green).

window. The performance is evaluated using the accuracy at frame level, which is defined as the number of correctetly estimated frames divided by the total number of frames. For that end, we have also included a new class 'no activity' associated to frames that are not showing any activity of interest. It is also worth noting that the global performance is computed by averaging the particular accuracies for each class (rather than simply counting the number of correct decisions) and, thus, adapts better to highly unbalanced sets as the one being used (where most of the time there is no activity of interest).

6.3.4.1 Window size

In our first experiment, we have studied the influence of the window size Δ defined in Sec. 6.2.3. Based on the results shown in Fig. 6.4 (blue line), we can draw interesting conclusions: on the one hand, too short windows do not model the dynamics of an activity, understood in our case as sequences of different active objects or places. Oppositely, too long windows may contain video segments showing various activities. Although, from our point of view, this fact might help to detect several strongly related activities by reinforcing the knowledge about one activity by the presence of the other (e.g. washing hands/face and drying hands/hair are activities that usually occur following the same temporal sequence), it might also lead to features containing too many active objects and places. These features would therefore make these frames difficult to assign to a particular activity. In our case, the value that best fits the activities in ADL dataset is $\Delta = 1200$ frames, which corresponds to approximately 47 secs of video footage. In fact, looking at the cumulative distribution of the activities length in the dataset (green line in Fig. 6.4), we have found this value is close to the median value which yields approximately 1100 frames, thereby being consistent with the intuition that the window size should be chosen to be representative of typical activities length.

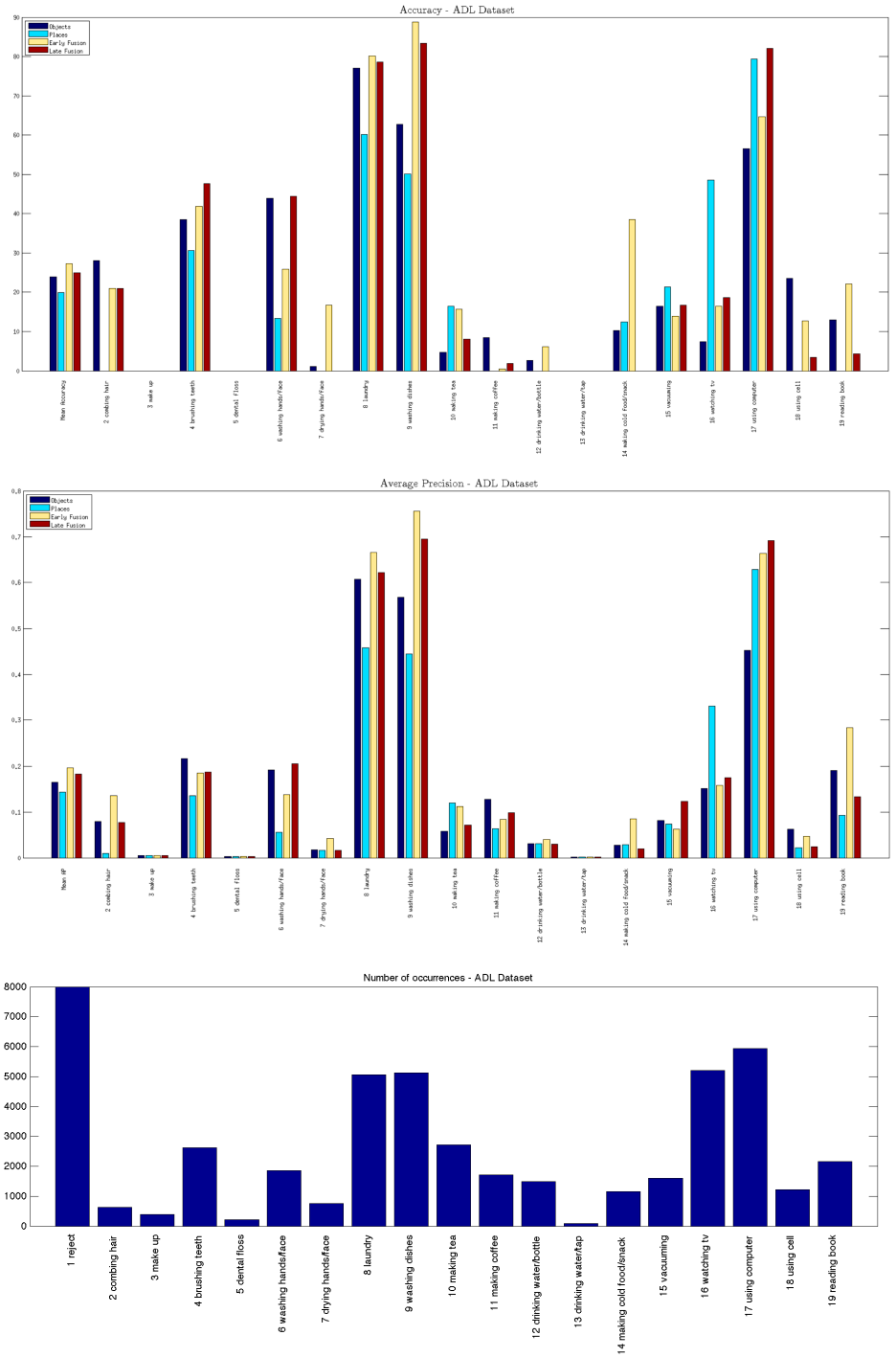


Figure 6.5 – (Top) Accuracy and (Middle) Average Precision for Activity Recognition for various strategies: Active Object alone, Places alone, early Fusion, late fusion of Active Objects and Places. (Bottom) Number of occurrences in the dataset.

Table 6.1 – Activity recognition accuracy for our approach computed at Frame and Segment level, respectively.

Approach	Avg Fr. Acc	Avg Seg. Acc
Active Objects	24.0%	37.4%
Places	18.5%	6.1%
Places + Rooms (early)	20.0%	11.1%
Active Objects + Places (early)	**27.3**%	38.5%
Active Objects + Places + Rooms (early)	26.3%	36.5%
Active Objects + Places (late)	25.0%	**40.0**%
Active Objects + Places + Rooms (late)	24.8%	39.3%
Pirsiavash et al. [Pirsiavash 12]	23.0%	36.9%

6.3.4.2 Recognition performance

In the first column of Table 6.1, we show the results of our approach using either just active object or place detectors, and using an early combination of both of them by feature concatenation. As one can notice from the results, the active objects using sliency alone achieves slightly better performance than the approach of [Pirsiavash 12]. The place and room information alone yield lower performance, possibly being less informative to discriminate the activities. Combining objects and their context (the place where they are located) notably improves the performance achieved by simply using the object detectors. Let us note that we have also tested several late fusion schemes (linear combinations, multiplicative, logarithmic, etc.) that did not lead to improvements in the system performance.

Furthermore, for comparison, we also include the results obtained with the software provided by the authors of [Pirsiavash 12]. This approach uses the outputs of various detectors of active and non-active objects implemented using the DPM from [Felzenszwalb 10]. Let us note that, as mentioned by the authors in the software, results differ from the ones reported in [Pirsiavash 12] due to changes in the dataset. From the results, and due to the similar classification pipeline of both methods, we can conclude that our features are more suitable for the activity recognition problem.

Finally, as made in [Pirsiavash 12], we additionally include results of a segment based evaluation in which ground truth time segmentations of the video are available in both training and testing steps. Hence, this case simplifies the activity recognition from a category segmentation problem to a simple classification problem for each segment. This case lacks the 'no activity' class, so that only video intervals showing activities of interest are taken into account. Combining objects and context provides the best performance, which is again superior to the one obtained by [Pirsiavash 12].

In order to analyze these results in more details, Figure 6.5 shows the Average Precision for each class separately. Performance is shown for several approaches, either using each mid-level feature alone, or using early or late fusion. It is clear from the results that sev-

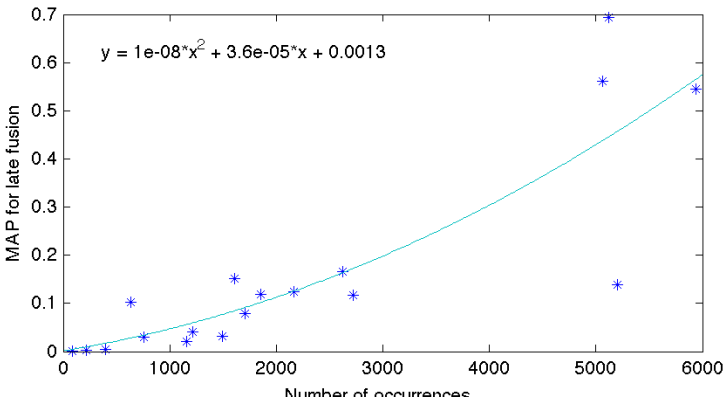


Figure 6.6 – mAP vs number of occurrences in dataset. Each star of the scatter plot represents one category. The best quadratic fitting is shown.

eral profiles appear for different kind of activities: some activities (Watching TV, Using the computer) are better recognized from object detection alone, while others (laundry, washing dishes, making coffee) are more linked to places. Their fusion tend to improve the mean performance, although the best way to do so depends on the activity category.

Overall, these results lead us to conclude that, recognizing activities in egocentric video does not require identifying every object in a scene, but simply detect the presence of 'active' objects and provide a compact representation of the object context. This context has been implemented in this work by means of a global classifier of the place. Future work could consider additional complementary features.

6.3.4.3 Amount of training data

It is very interesting to note that the performance seems to be positively correlated with the amount of training data available, as illustrated in Figure 6.6. There is indeed a sharp difference of average performance between the categories with a larger amount of training data and the others. Therefore, one main bottleneck of the recognition for this type of data remains the availability of sufficient training data, in order for the training to be representative of the test data. Although acquiring a large corpus of relevant data is an actual challenge when dealing with the monitoring of patients, these results suggest that ongoing and future efforts to obtain larger amount of training data in wearable camera setups is needed. Do they deal with control or patient subjects or not, they will likely contribute in notable improving the quality of the developped systems in terms of correct recognition of the activities.

6.4 Conclusion

In this chapter we have shown how activity recognition in egocentric video can be successfully addressed by the combination of two sources of information: a) active objects either manipulated or observed by the user provide very strong cues about the action, and b) con-

text also contributes with complementary information to the active objects, by identifying the place in which the action is being made.

For that end, an activity recognition method that models activities as sequences of actives objects and places have been used on a challenging egocentric video dataset showing daily living scenarios for various users. We have demonstrated how the combination of both objects+context provides notable improvements in the performance, and outperforms state-of-the-art methods using active+passive objects representations.

The results also show that activity recognition in unconstrained scenarios is still a challenging task, that requires the fusion of complementary sources of information. Future research directions may consider the use of additional complementary features such as motion, hand positions, presence of faces for social activities, and continue the very important task of collecting significant amount of wearable video data in order to improve the representativity of training datasets for the target tasks. This is actually the case in the first prototype of Dem@care system which is under tests with volunteers patients with dementia.

In the next and last chapter we will conclude this manuscript and present our work perspectives.

Chapter 7

Conclusions and perspectives

In this thesis, we studied the perceptual recognition of objects of interest. In this regard, we combined visual saliency - which models the selective process of human perception of visual scenes - with the Bag of Visual Words object recognition paradigm in order to focus the recognition towards active objects. We have presented different manner to integrate saliency into BoVW, from a trivial pooling approach to a deep integration at the heart of the vocabulary learning process. The studies we performed show how saliency improves the performances of object recognition, in particular the deep implementations. One important factor though is the quality of these saliency maps. Instead of relying of well-known saliency models, we proposed new ways to build visual attention maps fitted to the egocentric video content. We firstly studied how bottom-up models cannot necessarily rely on a central-bias hypothesis in this kind of content and introduced adapted replacements. Then we addressed the semantic top-down saliency approaches - less extensive because more challenging - by proposing a model specifically designed for egocentric content and showing great promises. All our models have been validated by psycho-visual experiments. In this work we aim at helping seniors and patients with dementia disease maintain their independence and stay at home longer, as well as helping medical practitioners in their studies of aging and age related dementia for the elaboration of adapted therapeutic treatments. Such are the objectives, in summary, of the Dem@care project under which this thesis was conducted. Under these objectives our goal was to help with the detection of IADLs of patients with dementia disease. Therefore our last proposed contribution consists in the study of IADL recognition for egocentric video content based on our perceptual object recognition techniques.

In this conclusion we come back on the different contributions introduced in this thesis and introduce various perspectives we believe relevant for the following of this work.

7.1 Summary of contributions

Pursuing our goal for IADL recognition, several contributions were introduced in the different domains listed below.

7.1.1 Visual Saliency

First of all, regarding visual saliency modeling we presented two contributions in both bottom-up and top-down domains.

In the bottom-up domain we have presented a method for building spatio-temporal saliency maps with a particular emphasis on geometrical cue in the case where a central-bias hypothesis does not hold. We proposed to combine spatial, temporal and geometrical cues as a linear combination with trained coefficients. The experiments have shown promising results as they highlight the necessity of a non-centered geometric saliency cue.

In the top-down domain we have proposed a probabilistic visual saliency model for the target task of recognition of manipulated objects in egocentric video. It is based on global and local features and uses domain knowledge, i.e. the fact that the object of interest is manipulated by hands. Experiments have shown how this top-down models outperforms several reference bottom-up models widely used in literature, both in terms of comparison with human gaze fixations and target performance in manipulated object recognition task. It improves the mean Average Precision scores by 13% over the standard BoVW model and outperforms other well-know models, establishing the new state-the-art for manipulated object recognition results in the considered dataset.

7.1.2 Active object recognition

Concerning the challenging active object recognition task in videos, we have proposed a perceptual model that incorporates visual attention. To do so, we have modeled independent foveal and peripheral pathways found in human retina, with particular properties in terms of spatial location, resolution, or sampling. In particular, we have introduced saliency into three particular processing modules of the well-known BoVW paradigm: (i) Visual Fields with Variable-Resolution and Non Uniform Sampling, (ii) Saliency-sensitive Coding of features, and (iii) Saliency-based Pooling. Rigorous experiments have shown how our biologically-inspired saliency-based model helps to enhance current system performance. It not only achieves notable improvements with respect to the baseline BoVW, but also provides state-of-the-art results in all the considered egocentric datasets at very competitive computational times. Furthermore, it avoids human efforts devoted to bounding-box level database annotation as in both training and test sets the saliency maps are automatically computed.

7.1.3 IADL recognition

Since our objective is to propose robust active object detectors for the IADL recognition task, our last contribution was to propose an activity recognition method that models activities

as sequences of actives objects and places. We tested our model on a challenging egocentric video dataset showing daily living scenarios for various users and have demonstrated how the combination of both objects+context provides notable improvements in the performance, and outperforms state-of-the-art methods using active+passive objects representations.

7.2 Limitations and Perspectives

This thesis opens various new perspectives that can be seen either as new directions or direct extension of this work. We also believe in complementary studies and further exploration of the available possibilities given by our contributions.

7.2.1 Deeper exploration of the models possibilities

This thesis was very vast in term of research possibilities, since visual saliency and object/action recognition are both huge fields of research of their own, and for that reason we were given the possibility to bring contributions to several domains as presented in 7.1. One of the drawbacks with such a sparse variety of possible contributions is sadly the lack of time to explore their full self or combined potentials. In this part we present what kind of studies we believe are still left to perform on our models.

We start by addressing the visual saliency field, and particularly our top-down probabilistic model. Indeed, we believe that this model could be extended to other domains of application and not only egocentric videos with detection of arms. The model could be adapted to many scenarios where there exist reference objects, which can be easily recognized, and to which the top-down attention can be related. One interesting example is the aided robotic surgery or the generation of post-surgery video reports. Here, the reference objects are the medical instruments, so that the attention is driven to the close operation field. Further examples are the recognition of e.g. robot-sorted objects on a conveyor belt, carried objects by a crane in a surveillance scenario. Another example, again with egocentric video content, is the real-time detection of objects with wearable glasses for manipulation by neuro-prostheses. Anyway this is a general principle: task-driven visual attention can be easily predicted if we can detect the presence of reference objects for such a task, which in this work where the hands of the user are performing the action.

About this top-down model, another track to explore is its integration with the new saliency-coding of features proposed in our object recognition contributions. Indeed with this integration model, we already obtain much improved recognition performances than the baseline BoVW using bottom-up saliency maps, and our top-down model outperforms greatly bottom-up models as seen in section 5.3.3. We believe the integration of both could lead to even greater results and it is one of our future perspectives of study.

We presented in this manuscript how a well fitted visual saliency model can focus the recognition towards the locus of objects. We believe an interesting contribution would be to use this information for the object localization task. Indeed, section 2.3.2.2 has listed the computational burden inherent to sliding window search. Visual saliency provides much

information about interesting locations and thus, if integrated properly, may greatly improve sliding-window techniques computation times.

Regarding activity recognition, we observed how challenging this task still remains, especially in unconstrained scenarios (that requires the fusion of complementary sources of information). Future research directions may consider the use of additional complementary features such as motion, hand positions, presence of faces for social activities, and continue the very important task of collecting significant amount of wearable video data in order to improve the representativity of training datasets for the target tasks. We also would like to explore how this model performs on different real-life datasets.

7.2.2 New combinations and models

We can consider new perspectives of research following our work that would lead to new studies.

First, in visual attention modeling we need to use domain knowledge and contextual information. Indeed, visual attention in the HVS is a complex combination of bottom-up, stimuli driven, and top-down, intentional components. In the perspective of the present research, the combination of bottom-up and top-down prediction in a video scene is envisaged with a target application to object and action recognition. One model could be to follow the process of the Human Visual System, that is to say relying on bottom-up features in the first, very fast unconscious step following a change of context, then giving more importance to the semantic information contained in the top-down maps.

Regarding the recognition of active objects, in this work we have seen how close we come to the best possible performances obtained with the classical BoVW paradigm (see section 4.2.5.4). In that sense we aim at exploring novel ways to introduce perceptual modeling into other object recognition models. It would be interesting indeed to see how visual attention modeling could improve the new best in class object recognition paradigms, such as deep learning architectures.

For the IADL recognition, we see several new tracks to explore. The first one would be to experiment with new saliency-based object recognition paradigms such as the one discussed before. Another would be to add our active object recognition paradigm in different models for activity recognition. One last contribution we see, using our model, would be to study how it could benefit from other sensor data, such as accelerometers for example.

7.2.3 Applications

In this last section we want to discuss the possible applications of this work. In our context of study, we have seen how IADL recognition can benefit from our models, and working closely with caregivers and doctors during this thesis proved the utility of our work for helping people with dementia disease such as Alzheimer's. Nowadays, with the democratization of wearable cameras (smartphones, egocentric cameras, embedded cameras,...), we believe this work could be extended to several other applications. Of course the first thing would be to develop a real-time implementation of our models. In the context of our study

this could lead to an on-line detection of activities of daily living of the patients, allowing to report to caregivers of doctors and get feedbacks right away. The fact that we aim at detecting objects of interest however could be used in many applications such as in personal robot assistants, drones applications, or other types of situations in which detecting active objects with an egocentric point of view is applicable and useful.

Chapter 8

Appendix - Detailed computation of the optimization of our top-down saliency model with Expectation-Maximization (see section 4.2.2.3)

In section 4.2.2.3 we have seen that the likelihood of the model parameters θ can be defined for the corpus (over all N pixels an D images considered) as:

$$\mathcal{L} = \prod_{d=1}^{D} p(\mathbf{z}_d) p(\mathbf{g}_d|\mathbf{z}_d) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|\mathbf{z}_d) \tag{8.1}$$

If we marginalize eq. 8.1 over the latent arm models we get a definition of the likelihood by means of a mixture of K components:

$$\mathcal{L} = \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \tag{8.2}$$

The log-likelihood is then given by:

$$\log \mathcal{L} = \sum_{d=1}^{D} \log \left(\sum_{k=1}^{K} p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \right) \tag{8.3}$$

We want to maximize this marginal log-likelihood. Expectation-Maximization is a very general algorithm for doing maximum likelihood estimation of parameters in models which

contain latent variables. Here $\mathbf{x}, \mathbf{g}$ are the observed variables, $\mathbf{z}$ the latent variables and θ the set of model parameters.

8.1 Defining a lower bound

We define the auxiliary distribution $\phi_{dk} = p(z_k|\mathbf{g}_d, \mathbf{x})$ to our model which stands for the posterior distribution of the arms model given the observed variables and obeys $\sum_k \phi_{dk} = 1$. Now let's re-write the complete log likelihood function by multiplying it by $\frac{\phi_{dk}}{\phi_{dk}}$. This gives:

$$\log \mathcal{L} = \sum_{d=1}^{D} \log \left(\sum_{k=1}^{K} p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \frac{\phi_{dk}}{\phi_{dk}} \right) \tag{8.4}$$

$$= \sum_{d=1}^{D} \log \left(\sum_{k=1}^{K} \phi_{dk} \frac{p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k)}{\phi_{dk}} \right) \tag{8.5}$$

$$= \sum_{d=1}^{D} \log \mathbb{E} \left(\frac{p(z) p(\mathbf{g}_d|z) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z)}{\phi_d} \right) \tag{8.6}$$

At this point we introduce Jensen's inequality. Jensen's Inequality states that for a convex function $f(x)$:

$$f(\mathbb{E}[x]) \leq \mathbb{E}\left[f(x)\right] \tag{8.7}$$

The reversed inequality holds for a concave function (such as $\log$), which we apply to lower bound the marginal log likelihood in eq. 8.6 :

$$\log \mathcal{L} = \sum_{d=1}^{D} \log \mathbb{E} \left(\frac{p(z) p(\mathbf{g}_d|z) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z)}{\phi_d} \right) \tag{8.8}$$

$$\geq \sum_{d=1}^{D} \mathbb{E} \left[\log \frac{p(z) p(\mathbf{g}_d|z) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z)}{\phi_d} \right] \tag{8.9}$$

$$\geq \sum_{d=1}^{D} \sum_{k=1}^{K} \phi_{dk} \left[\log \frac{p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k)}{\phi_{dk}} \right] \tag{8.10}$$

This gives us the lower bound in eq. 4.8:

$$\log \mathcal{L} \geq \sum_{d,k}^{D,K} \phi_{dk} \left[\log \left(p(z_k) p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \right) - \log \phi_{dk} \right] \tag{8.11}$$

8.2 Expectation step

The E-step consists of maximizing the lower bound defined in 8.11 with respect to ϕ_{dk}. This corresponds to maximizing the terms of the lower bound related to ϕ_{dk} defined by:

$$\mathcal{L}_{\phi_{dk}} = \phi_{dk}\left[\log\left(p(z_k)p(\mathbf{g}_d|z_k)\prod_{i=1}^{N}p(\mathbf{x}_i,\mathbf{c}_i,h_i|z_k)\right) - \log\phi_{dk}\right] - \lambda\left(\sum_{k=1}^{K}\phi_{dk} - 1\right) \tag{8.12}$$

where the term $\lambda\left(\sum_{k=1}^{K}\phi_{dk} - 1\right)$ is a Lagrange operator required because of the constraint $\sum_{k=1}^{K}\phi_{dk} = 1$.

Then in order to maximize this term with respect to ϕ_{dk} we must solve :

$$\frac{\partial\mathcal{L}_{\phi_{dk}}}{\partial\phi_{dk}} = 0 \tag{8.13}$$

$$\log\left(p(z_k)p(\mathbf{g}_d|z_k)\prod_{i=1}^{N}p(\mathbf{x}_i,\mathbf{c}_i,h_i|z_k)\right) - \log\phi_{dk} - 1 - \lambda = 0 \tag{8.14}$$

let us define $A_{dk} = p(z_k)p(\mathbf{g}_d|z_k)\prod_{i=1}^{N}p(\mathbf{x}_i,\mathbf{c}_i,h_i|z_k)$, we have:

$$\log A_{dk} - \log\phi_{dk} - 1 - \lambda = 0 \tag{8.15}$$

$$\log\frac{\phi_{dk}}{A_{dk}} = -\lambda - 1 \tag{8.16}$$

$$\phi_{dk} = A_{dk}e^{-\lambda-1} \tag{8.17}$$

From this we can derive:

$$\sum_{k=1}^{K}\phi_{dk} = \sum_{k=1}^{K}A_{dk}e^{-\lambda-1} \tag{8.18}$$

$$1 = \sum_{k=1}^{K}A_{dk}e^{-\lambda-1} \tag{8.19}$$

$$e^{-\lambda-1} = \frac{1}{\sum_{k=1}^{K}A_{dk}} \tag{8.20}$$

We can replace in eq. 8.17 to find:

$$\phi_{dk} = \frac{A_{dk}}{\sum_{k=1}^{K}A_{dk}} \tag{8.21}$$

$$\phi_{dk} \propto p(z_k)p(\mathbf{g}_d|z_k)\prod_{i=1}^{N}p(\mathbf{x}_i,\mathbf{c}_i,h_i|z_k) \tag{8.22}$$

8.3 Maximization step

The M-step then consists in maximizing the lower bound (eq. 8.11) with respect to the parameters $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$ and given the value for ϕ_{dk} computed in the E-step (eq. 8.22).

8.3.1 Considering the weights of the mixture: $p(z_k) = \pi_k$

Let us define the terms of the lower bound related to $p(z_k)$ by:

$$\mathcal{L}_{\pi_k} = \sum_d \phi_{dk} \log \pi_k - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{8.23}$$

where the term $\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$ is a Lagrange operator required because of the constraint $\sum_{k=1}^{K} \pi_k = 1$

Then in order to maximize this term with respect to π_k we must solve :

$$\frac{\partial \mathcal{L}_{\pi_k}}{\partial \pi_k} = 0 \tag{8.24}$$

$$\frac{1}{\pi_k} \sum_d \phi_{dk} - \lambda = 0 \tag{8.25}$$

$$\pi_k = \frac{\sum_d \phi_{dk}}{\lambda} \tag{8.26}$$

Since $\sum_{k=1}^{K} \pi_k = 1$ we have:

$$1 = \sum_k \frac{\sum_d \phi_{dk}}{\lambda} \tag{8.27}$$

$$1 = \frac{\sum_d \sum_k \phi_{dk}}{\lambda} \tag{8.28}$$

and since $\sum_{k=1}^{K} \phi_{dk} = 1$:

$$1 = \frac{D}{\lambda} \lambda = D \tag{8.29}$$

then from eq. 8.26:

$$\pi_k = \frac{1}{D} \sum_d \phi_{dk} \tag{8.30}$$

8.3.2 Considering the distributions of global features: $p(\mathbf{g}_d|z_k) = \mathcal{N}(\mathbf{g}; \mu_k^g, \Sigma_k^g)$

For the distribution of global features we have:

$$p(\mathbf{g}_d|z_k) = \mathcal{N}(\mathbf{g}_d; \mu_k^g, \Sigma_k^g) \tag{8.31}$$

$$= (2\pi)^{-\frac{6}{2}} |\Sigma_k^g|^{-\frac{1}{2}} exp(-\frac{1}{2}(g_d - \mu_k^g)^\top (\Sigma_k^g)^{-1}(g_d - \mu_k^g)) \tag{8.32}$$

Considering the means μ_k^g:

Let us define the terms of the lower bound related to the means μ_k^g by:

$$\mathcal{L}_{\mu_k^g} = \sum_d \phi_{dk} \left[-\frac{1}{2}(g_d - \mu_k^g)^\top (\Sigma_k^g)^{-1}(g_d - \mu_k^g) \right] \tag{8.33}$$

Then in order to maximize this term with respect to μ_k^g we must solve :

$$\frac{\partial \mathcal{L}_{\mu_k^g}}{\partial \mu_k^g} = 0 \tag{8.34}$$

$$-\sum_d \phi_{dk}(\Sigma_k^g)^{-1}(g_d - \mu_k^g) = 0 \tag{8.35}$$

$$-\sum_d \phi_{dk}(g_d - \mu_k^g) = 0 \tag{8.36}$$

$$\tag{8.37}$$

Hence:

$$\mu_k^g = \frac{\sum_d \phi_{dk} g_d}{\sum_d \phi_{dk}} \tag{8.38}$$

Considering the covariance matrices Σ_k^g:
Let us define the terms of the lower bound related to the covariance matrices Σ_k^g by:

$$\mathcal{L}_{\Sigma_k^g} = \sum_d \phi_{dk}\left[-\frac{1}{2}\log|\Sigma_k^g| - \frac{1}{2}(g_d - \mu_k^g)^\top(\Sigma_k^g)^{-1}(g_d - \mu_k^g)\right] \tag{8.39}$$

Then in order to maximize this term with respect to Σ_k^g we must solve :

$$\frac{\partial \mathcal{L}_{\Sigma_k^g}}{\partial \Sigma_k^g} = 0 \tag{8.40}$$

$$\sum_d \phi_{dk}\left[-\frac{1}{2}(\Sigma_k^g)^{-1} + \frac{1}{2}(\Sigma_k^g)^{-1}(g_d - \mu_k^g)(g_d - \mu_k^g)^\top(\Sigma_k^g)^{-1}\right] = 0 \tag{8.41}$$

Hence:

$$\Sigma_k^g = \frac{\sum_d \phi_{dk}(g_d - \mu_k^g)(g_d - \mu_k^g)^\top}{\sum_d \phi_{dk}} \tag{8.42}$$

8.3.3 Considering the distributions of local features: $p(\mathbf{x}, \mathbf{c}, h|z_k) = p(h|z_k)p(\mathbf{c}|h, z_k)p(\mathbf{x}|h, \mathbf{c}, z_k)$

From equation 4.4 we have seen that marginalizing over the variable h, we can expand the distribution involving the *local features*:

$$p(\mathbf{x}, \mathbf{c}, h|z_k) = \sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k) \tag{8.43}$$

So the terms of the lower bound depending on local features obey:

$$\mathcal{L}_{local} = \sum_{d=1}^{D}\sum_{k=1}^{K}\phi_{dk}\sum_{i=1}^{N}\log\left(\sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k)\right) \tag{8.44}$$

$$= \sum_{dki}\phi_{dk}\log\left(\sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k)\right) \tag{8.45}$$

Since maximizing $\mathcal{L}_{local}$ is non trivial, we employ Jensen's inequality to find a lower bound, followed by expectation maximization to maximize this lower bound similarly as before.

Applying Jensen's inequality (similarly to section 8.1) we obtain the lower bound:

$$\mathcal{L}_{local} \geq \sum_{dki} \phi_{dk} \sum_{j} \gamma_{dkij} \left[\log p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k) - \log \gamma_{dkij}\right] \tag{8.46}$$

where $\gamma_{dkij} = p(h_j|z_k,\mathbf{c},\mathbf{x}_i)$ is the posterior distribution of the selected hand once the global model, the center and the spatial location are known. γ_{dkij} obeys $\sum_{j=0}^{1} \gamma_{dkij} = 1$.

8.3.3.1 Expectation step

Here the E-step consists of maximizing the lower bound defined in 8.46 with respect to γ_{dkij}. This corresponds to maximizing the terms of the lower bound related to γ_{dkij} defined by:

$$\mathcal{L}_{\gamma_{dkij}} = \phi_{dk}\gamma_{dkij}\Bigg[\log p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k) - \log \gamma_{dkij}\Bigg] - \lambda\left(\sum_{j=0}^{1}\gamma_{dkij} - 1\right) \tag{8.47}$$

where the term $\lambda\left(\sum_{j=0}^{1}\gamma_{dkij} - 1\right)$ is a Lagrange operator required because of the constraint $\sum_{j=0}^{1}\gamma_{dkij} = 1$.

Then in order to maximize this term with respect to γ_{dkij} we must solve:

$$\frac{\partial \mathcal{L}_{\gamma_{dkij}}}{\partial \gamma_{dkij}} = 0 \tag{8.48}$$

$$\phi_{dk}\Bigg[\log p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k) - \log \gamma_{dkij} - 1\Bigg] - \lambda = 0 \tag{8.49}$$

Similarly to the computation for eq. 8.22, we can derive:

$$\gamma_{dkij} = \frac{p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k)}{\sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k)} \tag{8.50}$$

$$\gamma_{dkij} \propto p(h_j|z_k)p(\mathbf{c}|h_j,z_k)p(\mathbf{x}_i|h_j,\mathbf{c},z_k) \tag{8.51}$$

8.3.3.2 Maximization step

The M-step consists in maximizing the lower bound (eq. 8.46) with respect to the final parameters $\{\alpha, \mu^c, \Sigma^c, \beta\}$ and given the value for γ_{dkij} computed in the E-step (eq. 8.51).

- Considering $p(h_j|z_k) = \alpha_{jk}$

Let us define the terms of the lower bound in eq.8.46 related to α_{jk} by:

$$\mathcal{L}_{\alpha_{jk}} = \sum_d \sum_i \phi_{dk}\gamma_{dkij} . \log \alpha_{jk} - \lambda \left(\sum_{j=0}^{1} \alpha_{jk} - 1 \right) \tag{8.52}$$

where the term $\lambda \left(\sum_{j=0}^{1} \alpha_{jk} - 1 \right)$ is a Lagrange operator required because of the constraint $\sum_{j=0}^{1} \alpha_{jk} = 1$.

Then in order to maximize this term with respect to α_{jk} we must solve:

$$\frac{\partial \mathcal{L}_{\alpha_{jk}}}{\partial \alpha_{jk}} = 0 \tag{8.53}$$

$$\sum_d \sum_i \phi_{dk}\gamma_{dkij} \frac{1}{\alpha_{jk}} - \lambda = 0 \tag{8.54}$$

This gives us:

$$\alpha_{jk} = \frac{\sum_{di} \phi_{dk}\gamma_{dkij}}{\sum_{dij} \phi_{dk}\gamma_{dkij}} \tag{8.55}$$

- Considering the distribution for hand centers $p(\mathbf{c}|h_j, z_k) == \mathcal{N}(\mathbf{c}; \mu^c_{jk}, \Sigma^c_{jk})$

For the distribution of hand centers we have:

$$p(\mathbf{c}|h_j, z_k) == \mathcal{N}(\mathbf{c}; \mu^c_{jk}, \Sigma^c_{jk}) \tag{8.56}$$

$$= (2\pi)^{-\frac{2}{2}} |\Sigma^c_{jk}|^{-\frac{1}{2}} exp(-\frac{1}{2}(c_{dj} - \mu^c_{jk})^\top (\Sigma^c_{jk})^{-1}(c_{dj} - \mu^c_{jk})) \tag{8.57}$$

Considering the means μ^c_{jk}***:***

Let us define the terms of the lower bound in eq.8.46 related to the means mu^c_{jk} by:

$$\mathcal{L}_{\mu^c_{jk}} = \sum_{di} \phi_{dk}\gamma_{dkij} \left[-\frac{1}{2}(c_{dj} - \mu^c_{jk})^\top (\Sigma^c_{jk})^{-1}(c_{dj} - \mu^c_{jk}) \right] \tag{8.58}$$

Following the same computation process as in section 8.3.2 we derive:

$$\mu^c_{jk} = \frac{\sum_d \phi_{dk} c_{dj} \sum_i \gamma_{dkij}}{\sum_d \phi_{dk} \sum_i \gamma_{dkij}} \tag{8.59}$$

Considering the covariance matrices Σ^c_{jk}***:***

Let us define the terms of the lower bound in eq.8.46 related to the covariance matrices Σ^c_{jk} by:

$$\mathcal{L}_{\Sigma^c_{jk}} = \sum_{di} \phi_{dk}\gamma_{dkij} \left[-\frac{1}{2} \log |\Sigma^c_{jk}| - \frac{1}{2}(c_{dj} - \mu^c_{jk})^\top (\Sigma^c_{jk})^{-1}(c_{dj} - \mu^c_{jk}) \right] \tag{8.60}$$

Following the same computation process as in section 8.3.2 we derive:

$$\Sigma^c_{jk} = \frac{\sum_d \phi_{dk}(c_{dj} - \mu^c_{jk})(c_{dj} - \mu^c_{jk})^\top \sum_i \gamma_{dkij}}{\sum_d \phi_{dk} \sum_i \gamma_{dkij}} \tag{8.61}$$

- Considering the distribution $p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) = \beta_{kij}$
Let us define the terms of the lower bound in eq.8.46 related to β_{kij} by:

$$\mathcal{L}_{\beta_{kij}} = \sum_d \phi_{dk}\gamma_{dkij} \log \beta_{kij} - \lambda \left(\sum_{j=0}^{1} \beta_{kij} - 1 \right) \tag{8.62}$$

where the term $\lambda \left(\sum_{j=0}^{1} \beta_{kij} - 1 \right)$ is a Lagrange operator required because of the constraint $\sum_{i=0}^{N} \beta_{kij} = 1$.

Then in order to maximize this term with respect to β_{kij} we must solve:

$$\frac{\partial \mathcal{L}_{\beta_{kij}}}{\partial \beta_{kij}} = 0 \tag{8.63}$$

$$\sum_d \phi_{dk}\gamma_{dkij} \frac{1}{\beta_{kij}} - \lambda = 0 \tag{8.64}$$

This gives us:

$$\beta_{kij} = \frac{\sum_d \phi_{dk}\gamma_{dkij}}{\sum_d \phi_{dk} \sum_i \gamma_{dkij}} \tag{8.65}$$

8.4 Wrap-up

We have found a lower bound for the model and maximized it using Expectation maximization leading us to the following equations reported in section 4.2.2.3:

In the *E-Step*, the algorithm computes the expected values of the posterior distributions ϕ_{dk}, γ_{dkij}:

$$\phi_{dk} \propto p(z_k)p(\mathbf{g}_d|z_k) \prod_{i=1}^{N} p(x_i, \mathbf{c}_i, h_i|z_k) \tag{8.66}$$

$$\gamma_{dkij} \propto p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) \tag{8.67}$$

In the *M-Step*, our algorithm updates the values of the model parameters:

$$\pi_k = \frac{1}{D} \sum_d \phi_{dk} \tag{8.68}$$

$$\mu_k^g \propto \sum_d \phi_{dk} \mathbf{g}_d \tag{8.69}$$

$$\Sigma_k^g \propto \sum_d \phi_{dk} (\mathbf{g}_d - \mu_k^g)(\mathbf{g}_d - \mu_k^g)^T \tag{8.70}$$

$$\alpha_{jk} \propto \sum_{di} \phi_{dk} \gamma_{dkij} \tag{8.71}$$

$$\mu_{jk}^c \propto \sum_d \phi_{dk} c_{dj} \sum_i \gamma_{dkij} \tag{8.72}$$

$$\Sigma_{jk}^c \propto \sum_d \phi_{dk} (c_{dj} - \mu_{jk}^c)(c_{dj} - \mu_{jk}^c)^T \sum_i \gamma_{dkij} \tag{8.73}$$

$$\beta_{kji} \propto \sum_d \phi_{dk} \gamma_{dkij} \tag{8.74}$$

Bibliography

[Abdollahian 08] Golnaz Abdollahian, Zygmunt Pizlo & Edward J. Delp. *A study on the effect of camera motion on human visual attention*. In ICIP, pages 693–696. IEEE, 2008.

[Achanta 08] Radhakrishna Achanta, Francisco Estrada, Patricia Wils & Sabine SÃŒsstrunk. *Salient Region Detection and Segmentation*. In Antonios Gasteratos, Markus Vincze & JohnK. Tsotsos, editeurs, Computer Vision Systems, volume 5008 of *Lecture Notes in Computer Science*, pages 66–75. Springer Berlin Heidelberg, 2008.

[Ahonen 06] T. Ahonen, A. Hadid & M. Pietikainen. *Face Description with Local Binary Patterns: Application to Face Recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 12, pages 2037–2041, Dec 2006.

[Al 06] L. Shalabi Al & Z. Shaaban. *Normalization As a Preprocessing Engine for Data Mining and the Approach of Preference Matrix*. In Proceedings of the International Conference on Dependability of Computer Systems, DEPCOS-RELCOMEX '06, pages 207–214. IEEE Computer Society, 2006.

[Alexe 12] B. Alexe, T. Deselaers & V. Ferrari. *Measuring the Objectness of Image Windows*. IEEE Transactions on Pattern Analysis and Machine Intelligence,, vol. 34, no. 11, pages 2189–2202, 2012.

[Amieva 08] Hélène Amieva, Mélanie Le Goff, Xavier Millet, Jean Marc M. Orgogozo, Karine Pérès, Pascale Barberger-Gateau, Hélène Jacqmin-Gadda & Jean François F. Dartigues. *Prodromal Alzheimer's disease: successive emergence of the clinical symptoms.* Annals of neurology, vol. 64, no. 5, pages 492–498, November 2008.

[Andreopoulos 13] Alexander Andreopoulos & John K. Tsotsos. *50 Years of object recognition: Directions forward*. Computer Vision and Image Understanding, vol. 117, no. 8, pages 827 – 891, 2013.

[Avila 13] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle & Arnaldo de A. AraÃºjo. *Pooling in image representation: The visual codeword point of*

view. Computer Vision and Image Understanding, vol. 117, no. 5, pages 453 – 465, 2013.

[Baluja 94] Shumeet Baluja & Dean Pomerleau. *Using a Saliency Map for Active Spatial Selective Attention: Implementation & Initial Results*. In Advances in Neural Information Processing Systems, volume 6, pages 753–760, 1994.

[Bay 06] Herbert Bay, Tinne Tuytelaars & Luc Van Gool. *Surf: Speeded up robust features*. In In ECCV, pages 404–417, 2006.

[Bay 08] H. Bay, A. Ess, T. Tuytelaars & L. Van Gool. *Speeded-Up Robust Features (SURF)*. Comput. Vis. Image Underst., vol. 110, pages 346–359, June 2008.

[Belhumeur 96] PeterN. Belhumeur, JoÃ£oP. Hespanha & DavidJ. Kriegman. *Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection*. In Bernard Buxton & Roberto Cipolla, editeurs, Computer Vision - ECCV '96, volume 1064 of *Lecture Notes in Computer Science*, pages 43–58. Springer Berlin Heidelberg, 1996.

[Berry 07] Emma Berry, Narinder Kapur, Lyndsay Williams, Steve Hodges, Peter Watson, Gavin Smyth, James Srinivasan, Reg Smith, Barbara Wilson & Ken Wood. *The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report.* Neuropsychol Rehabil, vol. 17, no. 4-5, pages 582–601, Aug-Oct 2007.

[Betancourt 14] Alejandro Betancourt, Miriam M. Lopez, Carlo S. Regazzoni & Matthias Rauterberg. *A Sequential Classifier for Hand Detection in the Framework of Egocentric Vision*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2014.

[Biederman 95] Irving Biederman. Visual object recognition, volume 2. MIT press, 1995.

[Borji 12a] Ali Borji & Laurent Itti. *State-of-the-art in Visual Attention Modeling*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. PrePrints, 2012.

[Borji 12b] Ali Borji, DickyN. Sihite & Laurent Itti. *Salient Object Detection: A Benchmark*. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato & Cordelia Schmid, editeurs, Computer Vision - ECCV 2012, Lecture Notes in Computer Science, pages 414–429. Springer Berlin Heidelberg, 2012.

[Borji 13a] Ali Borji, Dicky N. Sihite & Laurent Itti. *Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study.* IEEE Transactions on Image Processing, vol. 22, no. 1, pages 55–69, 2013.

[Borji 13b] Ali Borji, Dicky N Sihite & Laurent Itti. *What stands out in a scene? A study of human explicit saliency judgment.* Vision Res, vol. 91, pages 62–77, Oct 2013.

[Borji 14] Ali Borji, Ming-Ming Cheng, Huaizu Jiang & Jia Li. *Salient Object Detection: A Survey*. CoRR, vol. abs/1411.5878, 2014.

[Boujut 12a] H. Boujut, J. Benois-Pineau & R. Megret. *Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion*. In ECCV 2012 - Workshops, ECCV'12, pages 436–445, 2012.

[Boujut 12b] Hugo Boujut. *Mesure sans référence de la qualité des vidéos haute définition diffusées avec des pertes de transmission*. PhD thesis, 2012. Thèse de doctorat dirigée par Benois Pineau, Jenny et Ahmed, Toufik Informatique Bordeaux 1 2012.

[Boujut 12c] Hugo Boujut, Jenny Benois-Pineau, Toufik Ahmed, Ofer Hadar & Patrick Bonnet. *No-Reference Video quality assessment of H.264 video streams based on semantic saliency maps*. In Image Quality and System Performance IX, volume 8293, pages 8293–28, San Francisco, États-Unis, January 2012. Contrat CIFRE avec Audemat Worldcast Systems.

[Boureau 10] Y.-L. Boureau, F. Bach, Y. LeCun & J. Ponce. *Learning mid-level features for recognition*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2559–2566, June 2010.

[Brouard 09] Olivier Brouard, Vincent Ricordel & Dominique Barba. *Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif*. In Compression et representation des signaux audiovisuels, CORESA 2009, page 6 pages, Toulouse, France, March 2009.

[Brown 07] Matthew Brown & DavidG. Lowe. *Automatic Panoramic Image Stitching using Invariant Features*. International Journal of Computer Vision, vol. 74, no. 1, pages 59–73, 2007.

[Bur 07] Alexandre Bur & Heinz Hügli. *Optimal Cue Combination for Saliency Computation: A Comparison with Human Vision*. In Proceedings of the 2nd international work-conference on Nature Inspired Problem-Solving Methods in Knowledge Engineering: Interplay Between Natural and Artificial Computation, Part II, IWINAC '07, pages 109–118, Berlin, Heidelberg, 2007. Springer-Verlag.

[Buswell 35] G. T. Buswell. How people look at pictures. The University of Chicago Press, Chicago, IL, USA, 1935.

[Cao 10] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang & Lei Zhang. *Spatial-bag-of-features*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 3352–3359, June 2010.

[Carrasco 11] Marisa Carrasco. *Visual attention: The past 25 years,*. Vision Research, vol. 51, no. 13, pages 1484–1525, July 2011.

[Cerf 07] M. Cerf, J. Harel, W. Einhäuser & C. Koch. *Predicting human gaze using low-level saliency combined with face detection.* In John C. Platt, Daphne Koller, Yoram Singer & Sam T. Roweis, editeurs, NIPS. Curran Associates, Inc., 2007.

[Chang 00] Ee-Chien Chang, Stéphane Mallat & Chee Yap. *Wavelet Foveation*. Applied and Computational Harmonic Analysis, vol. 9, no. 3, pages 312 – 335, 2000.

[Chang 11] Chih-Chung Chang & Chih-Jen Lin. *LIBSVM: A library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, vol. 2, pages 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chetverikov 02] Dmitry Chetverikov & Jirí Matas. *Periodic Textures as Distinguished Regions for Wide-Baseline Stereo Correspondence*, 2002.

[Chevalier 07] Fanny Chevalier, Jean-Philippe Domenger, Jenny Benois-Pineau & Maylis Delest. *Retrieval of objects in video by similarity based on graph matching*. Pattern Recognition Letters, vol. 28, no. 8, pages 939–949, 2007.

[Chum 09] O. Chum, M. Perdoch & J. Matas. *Geometric min-Hashing: Finding a (thick) needle in a haystack*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 17–24, June 2009.

[Cornelis 08] N. Cornelis & L. Van Gool. *Fast scale invariant feature detection and matching on programmable graphics hardware*. In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pages 1–8, June 2008.

[Cortes 95] C. Cortes & V. Vapnik. *Support-vector networks*. Machine Learning, vol. 20, pages 273–297, 1995.

[Csurka 04] G. Csurka, Ch. R. Dance, L. Fan, J. Willamowski & C. Bray. *Visual categorization with bags of keypoints*. In In Workshop on Statistical Learning in Computer Vision, ECCV, pages 1–22, 2004.

[Dalal 05] Navneet Dalal & Bill Triggs. *Histograms of Oriented Gradients for Human Detection*. In International Conference on Computer Vision & Pattern Recognition, volume 2, pages 886–893, June 2005.

[Dalal 06] Navneet Dalal, Bill Triggs & Cordelia Schmid. *Human Detection Using Oriented Histograms of Flow and Appearance*. In Aleš Leonardis, Horst Bischof & Axel Pinz, editeurs, Computer Vision – ECCV 2006, volume

3952 of *Lecture Notes in Computer Science*, pages 428–441. Springer Berlin Heidelberg, 2006.

[Daly 98] Scott J. Daly. *Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models*. In IS&T/SPIE Conference on Human Vision and Electronic Imaging III, volume 3299, pages 180–191, 1 1998.

[de Carvalho Soares 12] R. de Carvalho Soares, I.R. da Silva & D. Guliato. *Spatial Locality Weighting of Features Using Saliency Map with a BoVW Approach*. In International Conference on Tools with Artificial Intelligence, 2012, pages 1070–1075, 2012.

[Dean 13] Thomas Dean, Mark Ruzon, Mark Segal, Jonathon Shlens, Sudheendra Vijayanarasimhan & Jay Yagnik. *Fast, Accurate Detection of 100,000 Object Classes on a Single Machine*. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2013.

[Deng 09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li & L. Fei-Fei. *ImageNet: A Large-Scale Hierarchical Image Database*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, 2009.

[Detry 08] Renaud Detry, Nicolas Pugeault & Justus Piater. *Probabilistic Pose Recovery Using Learned Hierarchical Object Models*. In Barbara Caputo & Markus Vincze, editeurs, Cognitive Vision, volume 5329 of *Lecture Notes in Computer Science*, pages 107–120. Springer Berlin Heidelberg, 2008.

[Divvala 12] SantoshK. Divvala, AlexeiA. Efros & Martial Hebert. *How Important Are Deformable Parts in the Deformable Parts Model?* In Andrea Fusiello, Vittorio Murino & Rita Cucchiara, editeurs, Computer Vision - ECCV 2012. Workshops and Demonstrations, volume 7585 of *Lecture Notes in Computer Science*, pages 31–40. Springer Berlin Heidelberg, 2012.

[Dorr 10] M. Dorr, Th. Martinetz & E. Barth. *Variability of eye movements when viewing dynamic natural scenes*. Journal of Vision, vol. 10, no. 28, October 2010.

[Dovgalecs 13] Vladislavs Dovgalecs, Rémi Mégret & Yannick Berthoumieu. *Multiple Feature Fusion Based on Co-training Approach and Time Regularization for Place Classification in Wearable Video*. Adv. MultiMedia, vol. 2013, pages 1:1–1:22, January 2013.

[Duan 11] L. Duan, Ch. Wu & J. Miao. *Visual Conspicuity Index: Spatial Dissimilarity, Distance, and Central Bias*. IEEE Signal Processing Letters, vol. 18, no. 11, November 2011.

[Duda 72] Richard O. Duda & Peter E. Hart. *Use of the Hough Transformation to Detect Lines and Curves in Pictures*. Commun. ACM, vol. 15, no. 1, pages 11–15, January 1972.

[Elazary 08] Lior Elazary & Laurent Itti. *Interesting objects are visually salient*. J Vis, vol. 8, no. 3, pages 3.1–15, 2008.

[Erhan 14] Dumitru Erhan, Christian Szegedy, Alexander Toshev & Dragomir Anguelov. *Scalable Object Detection using Deep Neural Networks*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[Farnebäck 00] G. Farnebäck. *Fast and Accurate Motion Estimation using Orientation Tensors and Parametric Motion Models*. In Proceedings of 15th International Conference on Pattern Recognition, volume 1, pages 135–139, Barcelona, Spain, September 2000. IAPR.

[Fathi 11a] Alireza Fathi, Ali Farhadi & James M. Rehg. *Understanding egocentric activities*. In International Conference on Computer Vision, 2011, ICCV '11, pages 407–414, Washington, DC, USA, 2011.

[Fathi 11b] Alireza Fathi, Xiaofeng Ren & James M. Rehg. *Learning to recognize objects in egocentric activities*. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pages 3281–3288, 2011.

[Fathi 12] A. Fathi, Y. Li & J. M. Rehg. *Learning to Recognize Daily Actions Using Gaze*. In ECCV (1), pages 314–327, 2012.

[Felzenszwalb 10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester & D. Ramanan. *Object Detection with Discriminatively Trained Part-Based Models*. IEEE Transactions on Pattern Analyisis and Machine Intelligence, vol. 32, no. 9, pages 1627–1645, 2010.

[Fergus 03] R. Fergus, P. Perona & A. Zisserman. *Object class recognition by unsupervised scale-invariant learning*. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 2, pages II–264–II–271 vol.2, June 2003.

[Ferrari 06] Vittorio Ferrari, Tinne Tuytelaars & Luc Van Gool. *Simultaneous Object Recognition and Segmentation by Image Exploration*. In Jean Ponce, Martial Hebert, Cordelia Schmid & Andrew Zisserman, editeurs, Toward Category-Level Object Recognition, volume 4170 of *Lecture Notes in Computer Science*, pages 145–169. Springer Berlin Heidelberg, 2006.

[Fischler 73] Martin A. Fischler & R.A. Elschlager. *The Representation and Matching of Pictorial Structures*. Computers, IEEE Transactions on, vol. C-22, no. 1, pages 67–92, Jan 1973.

[Fischler 81] Martin A. Fischler & Robert C. Bolles. *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. Commun. ACM, vol. 24, pages 381–395, June 1981.

[Fleuret 01] Francois Fleuret & Donald Geman. *Coarse-to-Fine Face Detection*. International Journal of Computer Vision, vol. 41, no. 1-2, pages 85–107, 2001.

[Friedman 77] Jerome H. Friedman, Jon Louis Bentley & Raphael Ari Finkel. *An Algorithm for Finding Best Matches in Logarithmic Expected Time*. ACM Trans. Math. Softw., vol. 3, no. 3, pages 209–226, September 1977.

[Gaidon 09] Adrien Gaidon, Marcin Marszalek & Cordelia Schmid. *Mining visual actions from movies*. In A. Cavallaro and S. Prince and D. Alexander, editeur, British Machine Vision Conference, pages 125.1–125.11, Londres, United Kingdom, September 2009. British Machine Vision Association, BMVA Press. Page web de l'article : http://lear.inrialpes.fr/pubs/2009/GMS09/.

[Gao 09] D. Gao, S. Han & N. Vasconcelos. *Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 6, pages 989–1005, 2009.

[Gebel 09] M. Gebel. Multivariate calibration of classifier scores into probability space. VDM Publishing, Saarbrücken, Germany, 2009.

[Gionis 99] Aristides Gionis, Piotr Indyk & Rajeev Motwani. *Similarity Search in High Dimensions via Hashing*. In Proceedings of the 25th International Conference on Very Large Data Bases, VLDB '99, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[Girshick 13] Ross B. Girshick, Jeff Donahue, Trevor Darrell & Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. CoRR, vol. abs/1311.2524, 2013.

[Girshick 14] Ross B. Girshick, Forrest N. Iandola, Trevor Darrell & Jitendra Malik. *Deformable Part Models are Convolutional Neural Networks*. CoRR, vol. abs/1409.5403, 2014.

[Gold 96] Steven Gold & Anand Rangarajan. *A Graduated Assignment Algorithm for Graph Matching*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 4, pages 377–388, April 1996.

[González Díaz 13] I. González Díaz, V. Buso, J. Benois-Pineau, G. Bourmaud & R. Megret. *Modeling Instrumental Activities of Daily Living in Egocentric Vision As Sequences of Active Objects and Context for Alzheimer Disease Research*. In Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare, MIIRH '13, pages 11–14, New York, NY, USA, 2013. ACM.

[Grand-Brochier 11] Manuel Grand-Brochier. *Descripteurs 2D et 2D+t de points d'intérêt pour des appariements robustes*. Theses, Université Blaise Pascal - Clermont-Ferrand II, November 2011.

[Grauman 05] K. Grauman & T. Darrell. *The pyramid match kernel: discriminative classification with sets of image features*. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1458–1465 Vol. 2, Oct 2005.

[Grauman 11] Kristen Grauman & Bastian Leibe. Visual object recognition. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[Green 66] David M. Green & John A. Swets. Signal detection theory and psychophysics. Wiley, New York, 1966.

[Hachinski 94] V Hachinski. *Vascular dementia: a radical redefinition.* Dementia, vol. 5, no. 3-4, pages 130–132, May-Aug 1994.

[Harel 07] J. Harel, C. Koch & P. Perona. *Graph-based visual saliency*. Advances in neural information processing systems, vol. 19, page 545, 2007.

[Harris 88] Chris Harris & Mike Stephens. *A combined corner and edge detector*. In In Proc. of Fourth Alvey Vision Conference, pages 147–151, 1988.

[Helmer 06] Catherine Helmer, Karine Peres, Luc Letenneur, Luis Miguel Guttierez-Robledo, Hanta Ramaroson, Pascale Barberger-Gateau, Colette Fabrigoule, Jean-Marc Orgogozo & Jean-Francois Dartigues. *Dementia in subjects aged 75 years or over within the PAQUID cohort: prevalence and burden by severity.* Dement Geriatr Cogn Disord, vol. 22, no. 1, pages 87–94, 2006.

[Henderson] John M. Henderson. *Human gaze control during real-world scene perception*. Trends in Cognitive Sciences, vol. 7, no. 11, pages 498–504, 2015/08/26.

[Hodges 06] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur & Ken Wood. *SenseCam: A Retrospective Memory Aid*. In Proceedings of the 8th International Conference of Ubiquitous Computing (UbiComp 2006), pages 177–193. Springer Verlag, September 2006.

[Hoffman 95] James E. Hoffman & Baskaran Subramaniam. *The role of visual attention in saccadic eye movements*. Perception & Psychophysics, page 787795, 1995.

[Holzer 09] S. Holzer, S. Hinterstoisser, S. Ilic & N. Navab. *Distance transform templates for object detection and pose estimation*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1177–1184, June 2009.

[Hood 86] Donald C. Hood & Marcia A. Finkelstein. *Sensitivity to Light*. In K. R. Boff, L. Kaufman & J. P. Thomas, editeurs, Handbook of perception and human performance, Volume 1: Sensory processes and perception, chapitre 5, pages 5–1–5–66. John Wiley & Sons, New York, NY, 1986.

[Huawei 14] Tian Huawei, Fang Yuming, Zhao Yao, Lin Weisi, Ni Rongrong & Zhu Zhenfeng. *Salient region detection by fusion bottom-up and top-down features extracted from a single image*. IEEE Transcations on Image processing, vol. 23, no. 10, pages 4389–4398, May 2014.

[Indyk 98] Piotr Indyk & Rajeev Motwani. *Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality*. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[Itti 98] Laurent Itti, Christof Koch & Ernst Niebur. *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pages 1254–1259, November 1998.

[ITU 02] International Telecommunication Union ITU. *ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures*. Recommendation BT.500-11, International Telecommunication Union ITU, 2002.

[Ivanovici 10] Mihai Ivanovici, Noël Richard & Christine Fernandez-Maloigne. *Towards video quality metrics based on colour fractal geometry*. J. Image Video Process., vol. 2010, pages 4:1–4:18, January 2010.

[Jing 02] F. Jing, M. Li, H.J. Zhang & B. Zhang. *An effective region-based image retrieval framework*. In ACM International conference on Multimedia, 2002.

[Jones 99] Michael Jones, , Michael J. Jones & James M. Rehg. *Statistical Color Models with Application to Skin Detection*. In Computer Vision and Pattern Recognition, CVPR'1999, pages 274–280. IEEE Computer Society, 1999.

[Judd 09] Tilke Judd, Krista A. Ehinger, Frédo Durand & Antonio Torralba. *Learning to predict where humans look*. In ICCV, pages 2106–2113. IEEE, 2009.

[Kanan 09] C. Kanan, M. H. Tong, L. Zhang & G. W. Cottrell. *SUN: Top-down saliency using natural statistics*, 2009.

[Karaman 10] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J.-F. Dartigues & Y. Gaëstel. *Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases*. In International Conference on Pattern Recognition (ICPR), 2010, pages 4113 –4116, aug. 2010.

[Karaman 14] Svebor Karaman, Jenny Benois-Pineau, Vladislavs Dovgalecs, Rémi Mégret, Julien Pinquier, Régine André-Obrecht, Yann Gaëstel & Jean-François Dartigues. *Hierarchical Hidden Markov Model in detecting activities of daily living in wearable videos for studies of dementia*. Multimedia Tools and Applications, vol. 69, no. 3, pages 743–771, 2014.

[Kim 91] W.-Y. Kim & A.C. Kak. *3-D object recognition using bipartite matching embedded in discrete relaxation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 13, no. 3, pages 224–251, Mar 1991.

[Kitani 11] K.M. Kitani, T. Okabe, Y. Sato & A. Sugimoto. *Fast unsupervised ego-action learning for first-person sports videos*. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3241–3248, 2011.

[Koch 85] C. Koch & S. Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry.* Human neurobiology, vol. 4, no. 4, pages 219–227, 1985.

[Komogortsev 09] O. V. Komogortsev. *Gaze-contingent video compression with targeted gaze containment performance*. Journal of Electronic Imaging, vol. 18, no. 3, pages 033001–033001–10, 2009.

[Kraemer 04] P. Kraemer, Jenny Benois-Pineau & Jean-Philippe Domenger. *Scene Similarity Measure for Video Content Segmentation in the Framework of Rough Indexing Paradigm*. In Scene Similarity Measure for Video Content Segmentation in the Framework of Rough Indexing Paradigm, pages 141–155, Espagne, Aug 2004.

[Lampert 08] Christoph H. Lampert, Matthew B. Blaschko & Thomas Hofmann. *Beyond sliding windows: Object localization by efficient subwindow search*. In IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[Land 99] Michael Land, Neil Mennie & Jennifer Rusted. *The role of vision and eye movements in the control of activities of daily living*. Perception, vol. 28, pages 1311–1328, 1999.

[Laptev 08] I. Laptev, M. Marszalek, C. Schmid & B. Rozenfeld. *Learning realistic human actions from movies*. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8, June 2008.

[Larabi 09] Mohamed-Chaker Larabi, Pascal Pellegrin, Ghislain Anciaux, François-Olivier Devaux, Olivier Tulet, Benoît Macq & Christine Fernandez-Maloigne. *HVS-based Quantization Steps for Validation of Digital Cinema Extended Bitrates*. In SPIE, pages 1–10, États-Unis, 2009.

[Lazebnik 06] Svetlana Lazebnik, Cordelia Schmid & Jean Ponce. *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*.

In IEEE Conference on Computer Vision and Pattern Recognition - 2006, pages 2169–2178, 2006.

[Le Callet 01] Patrick Le Callet. *Critères objectifs avec référence de qualité visuelle des images couleur*. PhD thesis, Ecole Polytechnique de l'Université de Nantes, 2001.

[Le Meur 07] Olivier Le Meur, Patrick Le Callet & Dominique Barba. *Predicting visual fixations on video based on low-level visual features*. Vision Researchearch, vol. 47, no. 19, pages 2483–2498, Sep 2007.

[Lee 14] Stefan Lee, Sven Bambach, David J. Crandall, John M. Franchak & Chen Yu. *This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video*. June 2014.

[LeMeur 10] Olivier LeMeur, Alexandre Ninassi, Patrick Le Callet & Dominique Barba. *Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric*. Signal Processing Image Communication, vol. 25, no. 7, pages 547–558, 2010.

[LeMeur 12] O. LeMeur & T. Baccino. *Methods for comparing scanpaths and saliency maps: strengths and weaknesses.* Behav Res Methods, 2012.

[Lepetit 05] V. Lepetit, P. Lagger & P. Fua. *Randomized trees for real-time keypoint recognition*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 775–781 vol. 2, June 2005.

[Lewis 98] David D. Lewis. *Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval*. pages 4–15. Springer Verlag, 1998.

[Li 13a] Cheng Li & K.M. Kitani. *Model Recommendation with Virtual Probes for Egocentric Hand Detection*. In Computer Vision (ICCV), 2013 IEEE International Conference on, pages 2624–2631, Dec 2013.

[Li 13b] Cheng Li & K.M. Kitani. *Pixel-Level Hand Detection in Ego-centric Videos*. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3570–3577, June 2013.

[Li 14] Yin Li, Xiaodi Hou, C. Koch, J.M. Rehg & A.L. Yuille. *The Secrets of Salient Object Segmentation*. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 280–287, June 2014.

[Lindeberg 98] Tony Lindeberg. *Feature detection with automatic scale selection*. International Journal of Computer Vision, vol. 30, pages 79–116, 1998.

[Liu 07] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang & Heung-Yeung Shum. *Learning to Detect A Salient Object*. In Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1–8, June 2007.

[Liu 09] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. Doctoral Thesis, Massachusetts Institute of Technology, May 2009.

[Liu 11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang & H.Y. Shum. *Learning to detect a salient object*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 2, pages 353–367, 2011.

[Liversedge 11] S. Liversedge, I. Gilchrist & S. Everling. The oxford handbook of eye movements, chapter 33. Oxford Library of Psychology. OUP Oxford, 2011.

[Long 03] F. Long, H. Zhang & D.D. Feng. *Fundamentals of content-based image retrieval*. In Multimedia Information Retrieval and Management, 2003.

[Lowe 99] David G. Lowe. *Object Recognition from Local Scale-Invariant Features*. In Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[Lowe 04] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision, vol. 60, pages 91–110, 2004.

[Ma 05] Y. F. Ma, X. S. Hua, L. Lu & H. Zhang. *A generic framework of user attention model and its application in video summarization.* IEEE Transactions on Multimedia, vol. 7, no. 5, pages 907–919, 2005.

[MacQueen 67] J. MacQueen. *Some methods for classification and analysis of multivariate observations*. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[Mahadevan 13] V. Mahadevan & N. Vasconcelos. *Biologically Inspired Object Tracking Using Center-Surround Saliency Mechanisms*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 3, pages 541–554, 2013.

[Marat 09] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin & Anne Guérin-Dugué. *Modelling spatio-temporal saliency to predict gaze direction for short videos*. International Journal of Computer Vision, vol. 82, no. 3, pages 231–243, 2009. Département Images et Signal.

[Marszałek 06] Marcin Marszałek & Cordelia Schmid. *Spatial Weighting for Bag-of-Features*. In IEEE Conference on Computer Vision & Pattern Recognition, volume 2, pages 2118–2125, June 2006.

[Matas 02] J. Matas, O. Chum, M. Urban & T. Pajdla. *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions*. In Proc. BMVC, pages 36.1–36.10, 2002. doi:10.5244/C.16.36.

[Mathe 12] Stefan Mathe & Cristian Sminchisescu. *Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition*. In European Conference on Computer Vision (ECCV), 2012, pages 842–856, 2012.

[McKhann 84] G McKhann, D Drachman, M Folstein, R Katzman, D Price & E M Stadlan. *Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease.* Neurology, vol. 34, no. 7, pages 939–944, Jul 1984.

[Mégret 10] Rémi Mégret, Vladislavs Dovgalecs, Hazem Wannous, Svebor Karaman, Jenny Benois-Pineau, Elie El Khoury, Julien Pinquier, Philippe Joly, Régine André-Obrecht, Yann Gaëstel & Jean-François Dartigues. *The IMMED Project: Wearable Video Monitoring of People with Age Dementia*. In Proceedings of the International Conference on Multimedia, MM '10, pages 1299–1302, New York, NY, USA, 2010. ACM.

[Melloni 12] L. Melloni, S. Van Leeuwen, A. Alink & N. G. Muumlller. *Interaction between Bottom-up Saliency and Top-down Control: How Saliency Maps Are Created in the Human Brain.* Cereb Cortex, 2012.

[Meur 06] Olivier Le Meur, Patrick Le Callet, Dominique Barba & Dominique Thoreau. *A Coherent Computational Approach to Model Bottom-Up Visual Attention*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 5, pages 802–817, 2006.

[Mikolajczyk 04] Krystian Mikolajczyk & Cordelia Schmid. *Scale & Affine Invariant Interest Point Detectors*. International Journal of Computer Vision, vol. 60, no. 1, pages 63–86, 2004.

[Mikolajczyk 05] Krystian Mikolajczyk & Cordelia Schmid. *A performance evaluation of local descriptors*. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005.

[Mikolajczyk 06] K. Mikolajczyk, B. Leibe & B. Schiele. *Multiple Object Class Detection with a Generative Model*. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 1, pages 26–36, June 2006.

[Moosmann 06] Frank Moosmann, Diane Larlus & Frederic Jurie. *Learning saliency maps for object categorization*. In ECCV'06 Workshop on the Representation and Use of Prior Knowledge in Vision, 2006.

[Moravec 80] Hans Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. In tech. report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doctoral dissertation, Stanford University, numéro CMU-RI-TR-80-03. September 1980.

[Moreels 08] Pierre Moreels & Pietro Perona. *A Probabilistic Cascade of Detectors for Individual Object Recognition*. In Proceedings of the 10th European Conference on Computer Vision: Part III, ECCV '08, pages 426–439, Berlin, Heidelberg, 2008. Springer-Verlag.

[Mu 08] Yadong Mu, Shuicheng Yan, Yi Liu, T. Huang & Bingfeng Zhou. *Discriminative local binary patterns for human detection in personal album*. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8, June 2008.

[Nistér 06] David Nistér & Henrik Stewénius. *Scalable Recognition with a Vocabulary Tree*. In IN CVPR, pages 2161–2168, 2006.

[Ogaki 12] Keisuke Ogaki, Kris Makoto Kitani, Yusuke Sugano & Yoichi Sato. *Coupling eye-motion and ego-motion features for first-person activity recognition.* In IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2012, pages 1–7, 2012.

[Ojala 96] Timo Ojala, Matti Pietikäinen & David Harwood. *A comparative study of texture measures with classification based on featured distributions*. Pattern Recognition, vol. 29, no. 1, pages 51–59, January 1996.

[Olshausen 97] Bruno A. Olshausen & David J. Field. *Sparse coding with an overcomplete basis set: A strategy employed by V1?* Vision Research, vol. 37, no. 23, pages 3311 – 3325, 1997.

[Ott 11] P. Ott & M. Everingham. *Shared parts for deformable part-based models*. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1513–1520, June 2011.

[Ouerhani 03] Nabil Ouerhani, Roman von Wartburg, Heinz Hugli & Rene Muri. *Empirical Validation of the Saliency-based Model of Visual Attention*. Electronic Letters on Computer Vision and Image Analysis, vol. 3, no. 1, pages 13–24, 2003.

[Parkhurst 02] D. Parkhurst, K. Law & E. Niebur. *Modeling the role of salience in the allocation of overt visual attention*. Vision Research, vol. 42, pages 107–123, 2002.

[Peres 08] Karine Peres, Catherine Helmer, Helene Amieva, Jean-Marc Orgogozo, Isabelle Rouch, Jean-Francois Dartigues & Pascale Barberger-Gateau. *Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia: a prospective population-based study.* J Am Geriatr Soc, vol. 56, no. 1, pages 37–44, Jan 2008.

[Perronnin 07] F. Perronnin & C. Dance. *Fisher Kernels on Visual Vocabularies for Image Categorization*. In Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1–8, June 2007.

[Perry 02] Jeffrey S. Perry & Wilson S. Geisler. *Gaze-contingent real-time simulation of arbitrary visual fields*. In In Human Vision and Electronic Imaging, SPIE Proceedings, pages 57–69, 2002.

[Peters 05] Robert J. Peters, Asha Iyer, Laurent Itti & Christof Koch. *Components of bottom-up gaze allocation in natural images*. Vision Research, vol. 45, no. 18, pages 2397 – 2416, 2005.

[Peters 08] Robert J. Peters & Laurent Itti. *Applying Computational Tools to Predict Gaze Direction in Interactive Visual Environments*. ACM Trans. Appl. Percept., vol. 5, no. 2, pages 9:1–9:19, May 2008.

[Piccardi 07] L. Piccardi, B. Noris, O. Barbey, A. Billard, G. Schiavone, F. Keller & C. von Hofsten. *WearCam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children*. In Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, pages 594–598, Aug 2007.

[Pinto 13] Yair Pinto, Andries R. van der Leij, Ilja G. Sligte, Victor A. F. Lamme & H. Steven Scholte. *Bottom-up and top-down attention are independent*. Journal of Vision, vol. 13, no. 3, 2013.

[Pirsiavash 12] H. Pirsiavash & D. Ramanan. *Detecting Activities of Daily Living in First-person Camera Views*. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012.

[Platt 99] John C. Platt. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. In Advances in Large Margin Classifiers, pages 61–74. MIT Press, 1999.

[Pronobis 10] Andrzej Pronobis, Oscar M. Mozos, Barbara Caputo & Patric Jensfelt. *Multi-modal Semantic Place Classification*. The International Journal of Robotics Research (IJRR), vol. 29, no. 2-3, pages 298–320, February 2010.

[Rayner 98] Keith Rayner. *Eye movements in reading and information processing: 20 years of research*. Psychological Bulletin, pages 372–422, 1998.

[Ren 10] Xiaofeng Ren & Chunhui Gu. *Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video*. In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[Ren 14] Yi Ren, Aurélie Bugeau & Jenny Benois-Pineau. *Bag-of-bags of words - Irregular graph pyramids vs spatial pyramid matching for image retrieval*. In 4th International Conference on Image Processing Theory, Tools and Applications, IPTA 2014, Paris, France, October 14-17, 2014, pages 247–252, 2014.

[Riche 13] N. Riche, M. Duvinage, M. Mancas, B. Gosselin & T. Dutoit. *Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics*. In The IEEE International Conference on Computer Vision (ICCV), December 2013.

[Rothganger 06] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid & Jean Ponce. *3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints*. International Journal of Computer Vision, vol. 66, no. 3, pages 231–259, 2006.

[Rudoy 13] D. Rudoy, D. B. Goldman, E. Shechtman & L. Zelnik-Manor. *Learning Video Saliency from Human Gaze Using Candidate Selection*. In CVPR, pages 1147–1154. IEEE, 2013.

[Russell 08] BryanC. Russell, Antonio Torralba, KevinP. Murphy & WilliamT. Freeman. *LabelMe: A Database and Web-Based Tool for Image Annotation*. International Journal of Computer Vision, vol. 77, no. 1-3, pages 157–173, 2008.

[Salton 86] Gerard Salton & Michael J. McGill. Introduction to modern information retrieval. McGraw-Hill, Inc., New York, NY, USA, 1986.

[San Biagio 14] M. San Biagio, L. Bazzani, M. Cristani & V. Murino. *Weighted bag of visual words for object recognition*. In IEEE International Conference on Image Processing (ICIP), 2014, pages 2734–2738, Oct 2014.

[Schmid 97] Cordelia Schmid & Roger Mohr. *Local Grayvalue Invariants for Image Retrieval*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 5, pages 530–535, May 1997.

[Schmid 00] Cordelia Schmid, Roger Mohr & Christian Bauckhage. *Evaluation of Interest Point Detectors*. International Journal of Computer Vision, vol. 37, no. 2, pages 151–172, 2000.

[Schölkopf 98] Bernhard Schölkopf, Alexander Smola & Klaus-Robert Müller. *Nonlinear Component Analysis As a Kernel Eigenvalue Problem*. Neural Comput., vol. 10, no. 5, pages 1299–1319, July 1998.

[Schwartz 77] E.L. Schwartz. *Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception*. Biological Cybernetics, vol. 25, no. 4, pages 181–194, 1977.

[Seo 09] H.J. Seo & P. Milanfar. *Static and space-time visual saliency detection by self-resemblance*. Journal of Vision, vol. 9, no. 12, 2009.

[Sharma 12] G. Sharma, F. Jurie & C. Schmid. *Discriminative spatial saliency for image classification*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pages 3506–3513, 2012.

[Shen 14] Chengyao Shen & Qi Zhao. *Learning to Predict Eye Fixations for Semantic Contents Using Multi-layer Sparse Network*. Neurocomputing, vol. 138, pages 61–68, 2014.

[Shi 00] Jianbo Shi & Jitendra Malik. *Normalized Cuts and Image Segmentation*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pages 888–905, August 2000.

[Sirovich 87] L Sirovich & M Kirby. *Low-dimensional procedure for the characterization of human faces.* J Opt Soc Am A, vol. 4, no. 3, pages 519–524, Mar 1987.

[Sivic 03] J. Sivic & A. Zisserman. *Video Google: A Text Retrieval Approach to Object Matching in Videos*. In Proceedings of the International Conference on Computer Vision, volume 2, pages 1470–1477, October 2003.

[Sreekanth 10] V. Sreekanth, A. Vedaldi, C. V. Jawahar & A. Zisserman. *Generalized RBF feature maps for efficient detection*. In Proceedings of the British Machine Vision Conference (BMVC), 2010.

[Stoica 05] A. Stoica, M.C. Larabi & C. Fernez-Maloigne. *Visual quality enhancement for colour images in the framework of the JPEG2000 compression standard*. International Journal of Robotics and Automation 2005, vol. 20, no. 2, 2005.

[Su 14] Yingya Su, Qingjie Zhao, Liujun Zhao & Dongbing Gu. *Abrupt motion tracking using a visual saliency embedded particle filter*. Pattern Recognition, vol. 47, no. 5, pages 1826 – 1834, 2014.

[Sundaram 09] S. Sundaram & W.W.M. Cuevas. *High level activity recognition using low resolution wearable vision*. In CVPR Workshops 2009., pages 25–32, 2009.

[Swain 91] MichaelJ. Swain & DanaH. Ballard. *Color indexing*. International Journal of Computer Vision, vol. 7, no. 1, pages 11–32, 1991.

[Szegedy 15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke & Andrew Rabinovich. *Going Deeper with Convolutions*. In CVPR 2015, 2015.

[Tatler 05] Benjamin W. Tatler, Roland J. Baddeley & Iain D. Gilchrist. *Visual correlates of fixation selection: effects of scale and time*. Vision Research, vol. 45, no. 5, pages 643 – 659, 2005.

[Tatler 07] B. W. Tatler. *The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions*. Journal of Vision, vol. 7, no. 14, November 2007.

[Theriault 11] C. Theriault, N. Thome & M. Cord. *HMAX-S: Deep scale representation for biologically inspired image categorization*. In Image Processing (ICIP), 2011 18th IEEE International Conference on, pages 1261–1264, Sept 2011.

[Theriault 13] C. Theriault, N. Thome & M. Cord. *Extended Coding and Pooling in the HMAX Model*. Image Processing, IEEE Transactions on, vol. 22, no. 2, pages 764–777, Feb 2013.

[Tian 13] Yicong Tian, Rahul Sukthankar & Mubarak Shah. *Spatiotemporal Deformable Part Models for Action Detection*. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, pages 2642–2649, Washington, DC, USA, 2013. IEEE Computer Society.

[Tilke 12] J. Tilke, F. Durand & A. Torralba. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. 2012.

[Tong 11] Y. Tong, F. A. Cheikh, F. F. E. Guraya, H. Konik & A. Trémeau. *A Spatiotemporal Saliency Model for Video Surveillance.* Cognitive Computation, vol. 3, no. 1, pages 241–263, 2011.

[Torr 95] P. H. S. Torr & D. W. Murray. *Outlier Detection and Motion Segmentation*. In PhD Thesis, pages 432–443, 1995.

[Torralba 06] A. Torralba, M. S. Castelhano, A. Oliva & J. M. Henderson. *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search*. Psychological Review, vol. 113, page 2006, 2006.

[Treisman 80] Anne M. Treisman & Garry Gelade. *A feature-integration theory of attention*. Cognitive Psychology, vol. 12, no. 1, pages 97–136, January 1980.

[Tsotsos 95] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis & Fernando Nuflo. *Modeling visual attention via selective tuning*. Artificial Intelligence, vol. 78, no. 12, pages 507 – 545, 1995. Special Volume on Computer Vision.

[Turk 91] Matthew Turk & Alex Pentland. *Eigenfaces for Recognition*. J. Cognitive Neuroscience, vol. 3, no. 1, pages 71–86, January 1991.

[Tuytelaars 08] Tinne Tuytelaars & Krystian Mikolajczyk. *Local Invariant Feature Detectors: A Survey*. Found. Trends. Comput. Graph. Vis., vol. 3, no. 3, pages 177–280, July 2008.

[Uijlings 13] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers & A.W.M. Smeulders. *Selective Search for Object Recognition*. International Journal of Computer Vision, vol. 104, no. 2, pages 154–171, 2013.

[van de Sande 11] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers & Arnold W. M. Smeulders. *Segmentation As Selective Search for Object Recognition*. In Proceedings of the 2011 International Conference on Computer Vision, ICCV '11, pages 1879–1886, Washington, DC, USA, 2011. IEEE Computer Society.

[van Gemert 10] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders & J.-M. Geusebroek. *Visual Word Ambiguity*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 7, pages 1271–1283, July 2010.

[Vig 12] E. Vig, M. Dorr & D. D. Cox. *Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements*. In ECCV (7), pages 84–97, 2012.

[Viola 01] P. Viola & M. Jones. *Rapid object detection using a boosted cascade of simple features*. In IEEE Conference on Computer Vision and Pattern Recognition, volume 1, pages 511–518, April 2001.

[Viola 04] P. Viola & M.J. Jones. *Robust real-time face detection*. International journal of computer vision, vol. 57, no. 2, pages 137–154, 2004.

[Walther 04] Dirk Walther, Ueli Rutishauser, Christof Koch & Pietro Perona. *On the usefulness of attention for object recognition*. In Workshop on Attention and Performance in Computational Vision at ECCV, pages 96–103, 2004.

[Walther 06] Dirk Walther & Christof Koch. *Modeling attention to salient proto-objects*. Neural Networks, vol. 19, no. 9, pages 1395–1407, 2006.

[Wandell 95] B. A. Wandell. Foundations of vision. Sinauer Associates, Inc., 1995.

[Wang 10] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang & Yihong Gong. *Locality-constrained Linear Coding for Image Classification*. In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[Wannous 12] Hazem Wannous, Vladislavs Dovgalecs, Rémi Mégret & Mohamed Daoudi. *Place Recognition via 3D Modeling for Personal Activity Lifelog*

Using Wearable Camera. In Advances in Multimedia Modeling - 18th International Conference, MMM 2012, Klagenfurt, Austria, January 4-6, 2012. Proceedings, pages 244–254, 2012.

[Wei 13] Chia-Po Wei, Yu-Wei Chao, Yi-Ren Yeh & Yu-Chiang Frank Wang. *Locality-sensitive dictionary learning for sparse representation based classification*. Pattern Recognition, vol. 46, no. 5, pages 1277–1287, May 2013.

[Winn 05] J. Winn, A. Criminisi & T. Minka. *Object Categorization by Learned Universal Visual Dictionary*. In Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2, ICCV '05, pages 1800–1807, Washington, DC, USA, 2005. IEEE Computer Society.

[Wiskott 97] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger & Christoph Von Der Malsburg. *Face Recognition By Elastic Bunch Graph Matching*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 19, pages 775–779, 1997.

[Witkin 83] Andrew P. Witkin. *Scale-space Filtering*. In Proceedings of the Eighth International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'83, pages 1019–1022, San Francisco, CA, USA, 1983. Morgan Kaufmann Publishers Inc.

[Wong 07] Shu-Fai Wong & R. Cipolla. *Extracting Spatiotemporal Interest Points using Global Information*. In International Conference on Computer Vision, 2007. ICCV 2007., pages 1 –8, oct. 2007.

[Wooding 02] David Wooding. *Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps*. Behavior Research Methods, vol. 34, pages 518–528, 2002. 10.3758/BF03195481.

[Wu 04] Ting-Fan Wu, Chih-Jen Lin & Ruby C. Weng. *Probability Estimates for Multi-class Classification by Pairwise Coupling*. J. Mach. Learn. Res., vol. 5, pages 975–1005, December 2004.

[Yang 09] Jianchao Yang, Kai Yu, Yihong Gong & T. Huang. *Linear spatial pyramid matching using sparse coding for image classification*. In IEEE Conference on Computer Vision and Pattern Recognition, 2009., pages 1794–1801, 2009.

[Yarbus 67] Alfred Lukyanovich Yarbus. Eye movements and vision. Plenum. New York., 1967.

[Zhao 07] Guoying Zhao & M. Pietikainen. *Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 6, pages 915–928, June 2007.

[Zhong 13] Sh. Zhong, Y Liu, F. Ren, J. Zhang & T. Ren. *Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling*. In AAAI, 2013.

[Zhu 10] Long Zhu, Yuanhao Chen, A. Torralba, W. Freeman & A. Yuille. *Part and appearance sharing: Recursive Compositional Models for multi-view*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 1919–1926, June 2010.

Publications

International Journals

- V. Buso, I. Gonzalez Diaz, J. Benois-Pineau, *Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos,* Signal Processing: Image Communication, 2015; http://dx.doi.org/10.1016/j.image.2015.05.006
- V. Buso, J. Benois-Pineau, J.P. Domenger, *Geometrical cues in visual saliency models for active object recognition in egocentric videos,* Multimedia Tools and Applications, pp.1-19, 2015; http://dx.doi.org/10.1007/s11042-015-2803-2

International Conferences with Proceedings

- V. Buso, J. Benois-Pineau, I. Gonzalez Diaz, *Object Recognition with Top-Down Visual attention modeling for behavioral Studies,* ICIP, 2015; awarded best 10% papers.
- V. Buso, J. Benois-Pineau, P.M. Plans, L. Hopper, R. Mégret, *Recognition of Activities of Daily Living in Natural "At Home" Scenario for Assessment of Alzheimer disease patients,* Multimedia Expo Workshops (ICMEW), 2015 IEEE International Conference on, pp.1-6, Proceedings of the 1st International Workshop on Perception Inspired Video Processing, 2015; http://dx.doi.org/10.1109/ICMEW.2015.7169861
- V. Buso, J. Benois-Pineau, J.P. Domenger, *Geometrical Cues in Visual Saliency Models for Active Object Recognition in Egocentric Videos,* ACMMM PIVP '14, pp.9-14, Proceedings of the 1st International Workshop on Perception Inspired Video Processing, 2014; http://dx.doi.org/10.1145/2662996.2663007
- V. Buso, J. Benois-Pineau, I. Gonzalez Diaz, *Object recognition in egocentric videos with saliency-based non-uniform sampling and variable resolution space for features selection,* Workshop on Egocentric (First-person) Vision, 2014, CVPR

  I. Gonzalez Diaz, V. Buso, J. Benois-Pineau, G. Bourmaud, R. Megret, *Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for Alzheimer disease research,* ACMMM MIIRH '13, pp.11-14, Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, 2013; http://dx.doi.org/10.1145/2505323.2505328

H. Boujut, V. Buso, J. Benois-Pineau, Y. Gaestel, J.F. Dartigues, *Visual saliency maps for studies of behavior of patients with neurodegenerative diseases: Observer's versus Actor's points of view,* INMED, 2013;

Book Chapters

- I. Gonzalez-Diaz, V. Buso, J. Benois-Pineau, G. Bourmaud, G. Usseglio, R. Mégret, Y. Gaestel and J.F. Dartigues, *Recognition of Instrumental Activities of Daily Living in Egocentric Video for Activity Monitoring of patients with dementia,* Health Monitoring and Personalized Feedback using Multimedia Data, pp.161-178, Springer International Publishing, 2015; http://dx.doi.org/10.1007/978-3-319-17963-6_9
- I. Gonzalez-Diaz, V. Buso, H. Boujut, J. Benois-Pineau, *Fusion of Multiple Visual Cues for Object Recognition in Video,* Fusion in Computer Vision, pp.79-107, Advances in Computer Vision and Pattern Recognition, 2014, Springer International Publishing; http://dx.doi.org/10.1007/978-3-319-05696-8_4

National Conferences with Proceedings

- H. Boujut, V. Buso, G. Bourmaud, J. Benois-Pineau, R. Mégret, .P.-Domenger, Y. Gaestel, .F. Dartigues, *Egocentric vision IT technologies for Alzheimer disease assessment and studies,* RITS, 2013; http://arxiv.org/abs/1303.3134