INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité:
Imagerie Vision et Robotique

Ecole Doctoral:
Mathématiques, Sciences et technologies de l'information, Informatique

présentée et soutenue publiquement
par

Krystian Mikolajczyk

le 15 juillet 2002

Detection of local features invariant to affine transformations

Application to matching and recognition

Directeur de thèse: Cordelia Schmid

JURY

Roger Mohr, Président
Andrew Zisserman, Rapporteur
David Lowe, Rapporteur
Cordelia Schmid, Examinateur
Tony Lindeberg, Examinateur
Michel Dhome, Examinateur

Thèse préparée au laboratoire GRAVIR - IMAG au sein du projet MOVI
INRIA Rhône-Alpes, 655 av. de l'Europe, 38334 Sant Ismier, FRANCE

Abstract

In recent years the use of local characteristics has become one of the dominant approaches to content based object recognition. The detection of interest points is the first step in the process of matching or recognition. A local approach significantly improves and accelerates image retrieval from databases. Therefore a reliable algorithm for feature detection is crucial for many applications.

In this thesis we propose a novel approach for detecting characteristic points in an image. Our approach is invariant to geometric and photometric transformations, which frequently appear between scenes viewed in different conditions. We emphasize the problem of invariance to affine transformations. This transformation is particularly important as it can locally approximate the perspective deformations. Previous approaches provide partial solutions to this problem, as not all essential parameters of local features are estimated in an affine invariant way. Our method is truly invariant to affine transformations, which include significant scale changes.

An image is represented by a set of extracted points. The interest points are characterized by descriptors, which are computed with local derivatives of the neighborhoods of points. These descriptors together with a similarity measure enable point-to-point correspondences to be established, and as a result, the geometry between images to be computed. In the context of an image database, the descriptors are used to find similar points in the database, and therefore the similar image.

The usefulness of our method is confirmed by excellent results for matching and image retrieval. Several comparative evaluations show that our approach provided for larger progress in the context of these applications. In our experiments we use a large set of real images, enabling representative results to be obtained.

Keywords: Interest points, feature detection, affine invariance, scale invariance, feature description, matching, image retrieval, recognition.

Résumé

Une des approches dominantes pour la reconnaissance d'objets est basée sur les caractéristiques locales. La méthode utilise la description locale calculée au voisinage de points d'intérêt. La détection de points d'intérêt est une première étape dans le processus de la mise en correspondance et de la reconnaissance. L'approche par apparences locales a permis d'améliorer et d'accélérer considérablement la recherche d'images dans des bases de données.

Dans cette thèse, nous proposons une nouvelle approche pour la détection de points caractéristiques d'une image. Cette approche est invariante aux transformations géométriques et photométriques, qui apparaissent fréquemment entre les images prises dans des conditions différentes. Nous nous concentrons sur le problème d'invariance aux transformations affines. Cette transformation est particulièrement importante parce qu'elle permet de s'affranchir des problèmes de changements perspectives. Les approches précédentes apportent des solutions partielles, car certains paramètres de points d'intérêt ne sont pas estimés de façon invariante aux changements affines. Nous avons proposé une solution générique à ces problèmes. Notre méthode est réellement invariante aux transformations affines, y compris aux changements d'échelle importants.

Les images sont caractérisées par des ensembles de descripteurs calculés en des points caractéristiques détectés automatiquement. Une mesure de ressemblance permet d'établir des correspondances entre les points. Ces correspondances sont ensuite utilisées pour calculer la géométrie qui lie les images. Dans le contexte de la recherche d'images les descripteurs sont utilisés pour retrouver des points similaires dans la base et par conséquent des images similaires aux images requêtes.

Les résultats expérimentaux pour la mise en correspondance et la recherche d'images montrent que notre approche est très robuste et efficace même dans les cas de changements importants. Plusieurs études comparatives effectuées dans cette thèse montrent l'avantage de cette méthode par rapport aux approches existantes présentées récemment dans la littérature.

Mots Cles: Points d'intérêt, détection de points caractéristiques, invariance affine, description locale, mise en correspondance, recherche d'images, reconnaissance par apparence locale.

Acknowledgement

I would like to thank all those people who have contributed to this work.

First of all I would to thank my advisor, Cordelia Schmid, for her guidance in this research.

I would like to thank Roger Mohr who invited me to prepare the thesis in the MOVI team.

I am grateful to David Lowe and Andrew Zisserman with whom I had valuable discussions, provided feedback on my research and helpful suggestions.

In my thanks I do not forget about Tony Lindeberg, whose work was the main source of inspiration during my research.

Finally, thanks to all former and actual MOVI members for an excellent ambiance during work.

Most of all I would like to thank Agnieszka, who has been sharing with me the happy and the difficult moments during last three years.

Contents

1 Résumé de la thèse **13**
1.1 Objectives et approches . . . 13
1.2 Contenu de la thèse . . . 16
1.2.1 Théorie de caractéristiques locales . . . 16
1.2.2 Détecteurs de points d'intérêt . . . 17
1.2.3 Description locale de points d'intérêts . . . 18
1.2.4 Appariement et reconnaissance d'images . . . 19
1.2.5 Reconnaissance de classes d'objets . . . 20
1.3 Conclusion et perspectives . . . 22

2 Introduction **25**
2.1 Principal issues . . . 25
2.1.1 Matching and recognition of objects . . . 26
2.1.2 Recognition of an object class . . . 28
2.2 Contributions . . . 30
2.3 Overview . . . 31

3 Theory of local features **33**
3.1 Multi-scale representation . . . 33
3.1.1 Gaussian scale-space . . . 34
3.1.2 Scale-space derivatives . . . 37
3.1.3 Second moment matrix . . . 40
3.1.4 Hessian matrix . . . 43
3.2 Automatic scale selection . . . 43
3.2.1 Scale-space maxima . . . 44
3.2.2 Gamma normalization . . . 46
3.2.3 Differential expressions for scale selection . . . 47
3.2.4 Experimental evaluation . . . 48
3.3 Discussion . . . 54

4 Interest point detectors **55**
4.1 Scale invariant detector . . . 55
4.1.1 State of the art . . . 56
4.1.2 Harris-Laplace detector . . . 58

4.1.3 Scale covariant points . . . 59
4.2 Affine invariant detector . . . 61
4.2.1 State of the art . . . 61
4.2.2 Harris-Affine detector . . . 63
4.2.3 Affine covariant points . . . 68
4.3 Comparative evaluation of detectors . . . 69
4.3.1 Scale invariant detectors . . . 70
4.3.2 Affine invariant detectors . . . 71
4.4 Discussion . . . 72

5 Local image description **75**
5.1 State of the art . . . 76
5.2 Differential description . . . 78
5.2.1 Differential invariants . . . 78
5.2.2 Dominant orientation estimation . . . 80
5.2.3 Noise . . . 82
5.2.4 Distance measure . . . 83
5.2.5 Variance and covariance of differential invariants . . . 84
5.3 Entropy of descriptor space . . . 87
5.3.1 Entropy criterion . . . 87
5.3.2 Results for information content . . . 87
5.4 Discussion . . . 88

6 Matching and indexing **91**
6.1 Matching and indexing algorithm . . . 91
6.1.1 Detection of scale covariant points . . . 92
6.1.2 Detection of affine covariant points . . . 92
6.1.3 Description . . . 93
6.1.4 Similarity measure . . . 93
6.1.5 Robust matching . . . 93
6.1.6 Retrieval algorithm . . . 93
6.2 Experimental results for matching . . . 94
6.2.1 Scale change . . . 94
6.2.2 Significant viewpoint change . . . 95
6.3 Experimental results for image retrieval . . . 95
6.3.1 Scale change problem . . . 97
6.3.2 Perspective deformation . . . 101
6.4 Discussion . . . 101

7 Recognition of an object class **105**
7.1 Introduction . . . 105
7.1.1 Motivations . . . 106
7.1.2 Related work . . . 107
7.2 Face detector . . . 109

7.2.1 Introduction to wavelets . . . 109
7.2.2 Appearance representation . . . 112
7.2.3 Probability score . . . 114
7.2.4 Pose representation . . . 115
7.2.5 Detection algorithm . . . 117
7.3 Face detection in a video sequence . . . 117
7.3.1 Temporal approach . . . 117
7.3.2 Adapted condensation algorithm . . . 118
7.4 Experimental results . . . 121
7.4.1 Single frame detection . . . 121
7.4.2 Temporal detection . . . 122
7.4.3 Observations . . . 122
7.5 Discussion . . . 123

8 Discussion **131**
8.1 Conclusions . . . 131
8.2 Future work . . . 133
8.2.1 Matching and recognition of rigid objects . . . 134
8.2.2 Recognition of an object class . . . 135

A ANNEX **139**
A.1 Extremum of local derivatives . . . 140
A.2 Repeatability criterion . . . 141
A.3 Test images . . . 143

Chapter 1

Résumé de la thèse

LA RECONNAISSANCE des formes dans les images est un problème central de l'intelligence artificielle. La quantité abondante de données numériques nécessite des outils automatiques et performants pour la reconnaissance et la recherche de l'information pertinente. L'objectif principal de la reconnaissance des formes est d'extraire le contenu sémantique des images. Pour représenter le contenu d'images on peut analyser les structures locales et ensuite utiliser les structures les plus discriminantes. La reconnaissance d'objets à partir de points d'intérêt a été initiée par Schmid et Mohr [107]. Cette approche s'est avérée particulièrement réussie et performante. Les points d'intérêt ont permis d'accélérer et d'améliorer considérablement la reconnaissance de formes dans les images.

1.1 Objectives et approches

Dans cette thèse, nous abordons le problème d'extraction de points caractéristiques d'images qui sont ensuite utilisés pour apparier les images. L'appariement consiste à identifier des parties communes à deux images. La deuxième partie de la recherche s'inscrit dans le domaine de la reconnaissance de classes d'objets.

Appariement d'images. L'objectif principal de la thèse est de développer une méthode de détection de caractéristiques locales, pour permettre la reconnaissance et la mise en correspondance invariante aux changements géométriques importants. Les approches de reconnaissance et de mise en correspondance sont basées sur les points d'intérêts mais les objectifs et les techniques employées sont différents. La détection de points caractéristiques est la première étape dans chacune de ces approches et les résultats finaux dépendent fortement de la stabilité et de la quantité des points extraits. Dans cette thèse nous allons donc nous concentrer sur la détection de points d'intérêt. Nous suivons le schéma classique de l'approche locale. La première étape consiste à détecter dans une image un ensemble de points d'intérêt. L'objectif est d'extraire les points d'une façon invariante à toutes les transformations, dues aux conditions arbitraires de la prise de vue. Ensuite, les points sont

caractérisés par des grandeurs numériques. Une mesure de ressemblance permet de calculer la similarité entre les points, afin de les apparier. Dans le cas de la mise en correspondance, les descripteurs sont utilisés pour établir des paires de points correspondants et ensuite pour calculer la géométrie entre les images. Dans le contexte de la reconnaissance, les descripteurs permettent de trouver un modèle dans une base d'images, qui est le plus ressemblant à l'image analysée.

Un des problèmes principaux est d'extraire des caractéristiques locales représentatives pour l'image et invariantes aux changements arbitraires de conditions de prises de vue. Les images réelles, en général, présentent des scènes avec des objets partiellement lisses. Une surface lisse peut être approximée par des surfaces planes. La surface plane est déformée par la transformation perspective si elle est vue sous des angles différents. Finalement, la transformation perspective peut être localement approximée par une transformation affine. Par conséquent, nous constatons que les caractéristiques locales invariantes aux changements affines permettront la reconnaissance d'objets partiellement lisses sans aucune connaissance à priori sur les conditions de vue. Le problème le plus important à résoudre est l'influence de la transformation affine sur les paramètres principaux, qui décrivent l'apparence d'une structure locale, c'est-à-dire la localisation, la taille, la rotation et la forme. Il est alors essentiel d'estimer la déformation affine de la région, avant de déterminer les autres paramètres. Pour remédier à ce problème, on utilise les valeurs propres de la matrice de deuxième moment. La taille de la région est détectée par la technique de sélection automatique d'échelle. L'estimation de l'orientation dominante liée à la structure locale permet le calcul des descripteurs invariants à la rotation.

Reconnaissance de classes d'objets. La deuxième partie de la thèse est consacrée à la reconnaissance de classes d'objets. La détection de visages est un excellent exemple de la classification. Un algorithme efficace pour la détection de visage trouve ses applications dans les nombreux domaines de la vision par ordinateur. L'objectif de la détection est de déterminer la localisation, la taille et la pose du visage dans une image. Le visage humain est une structure relativement discriminante mais simultanément complexe dans son apparence locale et globale. Le problème de la reconnaissance de classes d'objets est différent de la reconnaissance d'un même objet ou d'une scène vue de points de vue différents. L'approche proposée pour l'appariement ne donne pas de résultats satisfaisants dans le cas de la classification. L'objet à reconnaître peut prendre des apparences différentes, par exemple couleurs, pose, expression etc. Un model d'apparence doit capturer toutes ces variations. Nous avons adapté l'algorithme de classification proposé par Schneiderman et Kanade [110]. Cette approche est basée sur deux modèles: un pour les visages et l'autre pour les objets différents du visage. Le modèle du visage représente la distribution statistique des apparences du visage. Les apparences locales sont codées par les coefficients d'ondelette et accumulées dans des histogrammes multidimensionnels. Une image requête est comparée avec le modèle du visage et ensuite avec le modèle du non-visage. Une mesure de similarité permet de classifier l'image requête selon le résultat de cette comparaison.

La séquence vidéo contient une information temporelle qui peut être utilisée pour stabiliser les résultats de détection et pour éliminer les fausses alarmes. Nous utilisons cette information dans la phase de la prédiction et de la mise à jour. Ainsi on limite la

zone de recherche dans l'image et accélère considérablement les calculs. Un des problèmes difficiles dans la détection de visages est le changement de pose du visage. La baisse de probabilité d'occurrence d'un visage au cours du temps dans la vidéo peut être aussi bien due aux changements de pose qu'aux occultations. Pour résoudre ce problème nous proposons d'utiliser deux détecteurs, un pour les visages vus de face et l'autre pour les profils, ainsi que l'information temporelle dans la séquence. La combinaison de ces deux détecteurs permet de détecter aussi les poses intermédiaires.

Contributions principales. Dans le contexte d'appariement d'images à des échelles différentes nous avons proposé d'utiliser la sélection automatique d'échelle pour déterminer la taille du voisinage d'un point d'intérêt. Nous avons mené une étude théorique et une évaluation expérimentale de la technique de sélection automatique d'échelle. Nous avons testé plusieurs expressions différentielles et déterminé les plus adaptées à notre problème.

Nous avons proposé une nouvelle méthode de détection de points d'intérêt invariante aux changements d'échelle. Cette approche combine la mesure de Harris [48] et celle du Laplacien [70] et permet d'obtenir des résultats meilleurs que chacune des méthodes appliquées séparément. Les points sont très discriminants et représentatifs pour l'image.

La contribution principale de la thèse est le détecteur de points invariant à la transformation affine. Nous avons introduit une mesure d'isotropie pour détecter la déformation affine d'une structure locale isotrope. Dans les solutions existantes, certains paramètres importants ne sont pas estimés d'une façon invariante aux changements affines. Notre approche estime simultanément tous les paramètres qui sont affectés par la transformation affine et détecte les mêmes points avec des régions associées dans des images déformées par une transformation projective.

Dans le contexte d'invariance à la rotation nous avons proposé une méthode stable d'estimation de l'orientation dominante du gradient dans le voisinage d'un point d'intérêt. Ceci permet d'effectuer une rotation préliminaire du voisinage et d'obtenir l'invariance à la rotation pour un descripteur quelconque.

Une des contributions de ce travail est l'analyse théorique et expérimentale de descripteurs différentiels. Nous avons montré que ces descripteurs sont sensibles aux différents types du bruit et qu'il existe une corrélation entre leurs composants. Nous proposons une méthode d'apprentissage de la mesure de similarité et d'estimation de la corrélation entre les invariants différentiels.

Pour évaluer et comparer différentes méthodes de détections nous avons proposé un ensemble de critères qui prennent en compte la localisation, l'échelle, la déformation affine et le caractère discriminant d'un point d'intérêt. Ces critères sont utilisés pour évaluer la performance des détecteurs.

Nous avons validé nos approches dans le contexte de la mise en correspondance et de la reconnaissance d'images. Les tests ont été menés sur une grande quantité de données pour obtenir des résultats fiables. Les résultats des tests montrent une excellente performance de nos approches.

Le détecteur de visage que nous avons développé, intègre les avantages de la détection dans des images fixes et l'aspect temporel d'une vidéo. Ceci permet d'améliorer la détection et d'effectuer le suivi de visages dans des séquences vidéo. La combinaison de deux

détecteurs, un pour les vues de face et l'autre pour les profils, permet de s'affranchir de problèmes de changements de pose.

1.2 Contenu de la thèse

Nous présentons par la suite le contenu de la thèse. Chaque section correspond à un chapitre de ce manuscrit et dans chaque paragraphe nous décrivons brièvement le contenu de la section correspondante dans la partie anglaise.

1.2.1 Théorie de caractéristiques locales

Le chapitre 3 présente la théorie sur laquelle sont fondées les algorithmes de détection de points d'intérêt proposés dans cette thèse.

La représentation multi-échelle. La section 3.1 introduit le cadre multi-échelle. La représentation en échelle est un ensemble d'images qui représentent une scène à plusieurs niveaux de résolution. Les niveaux de résolution sont calculés par une fonction adéquate paramétrée par un facteur d'échelle. La fonction doit être normalisée par rapport à l'échelle. Des nombreux résultats de recherches ont montré que la Gaussienne est optimale pour construire l'espace d'échelle. Les caractéristiques locales sont détectes par des fonctions qui combinent les dérivées gaussiennes d'ordres différents. La matrice du deuxième moment et la Hessienne sont constituées des dérivées gaussiennes et sont souvent exploitées dans le contexte de la détection de points d'intérêt. Les propriétés de la matrice du deuxième moment permettent d'aborder le problème de transformation affine du voisinage d'un point. La déformation affine d'une région centrée sur le point d'intérêt est estimée à partir des valeurs propres de cette matrice. La trace et le déterminant de la Hessienne sont particulièrement bien adaptés pour la détection automatique de l'échelle caractéristique d'un point. Ceci nous sera utile dans la suite pour la détection de points invariants aux changements d'échelle.

Détection automatique d'échelle. Les propriétés d'échelles caractéristiques sont étudiées dans la section 3.2. Nous présentons la technique de la sélection d'échelle associée à la structure locale qui sera ensuite utilisée dans notre approche pour la détection de points d'intérêt. Les points caractéristiques sont indiqués par des maxima locaux dans l'espace d'échelle construit à partir de dérivées normalisées. Dans ce contexte plusieurs fonctions qui combinent les dérivées normalisées ont été proposées dans la littérature. Les fonctions dépendent du type d'information que l'on veut extraire d'une image (i.e. régions homogènes, contours, coins). Pour un point donné on calcule des réponses d'une telle fonction pour plusieurs facteurs d'échelle. Ainsi, on construit la représentation en échelle pour le voisinage de ce point. Dans l'ensemble des réponses, on cherche une échelle caractéristique relativement indépendante de la résolution de l'image. Cette échelle caractéristique est indiquée par un maximum local des réponses. Le quotient des échelles trouvées indépendamment pour deux points correspondants est égal au facteur d'échelle réel entre les deux images. L'évaluation comparative des fonctions différentes menée dans

cette section montre que la mesure du Laplacien permet d'obtenir les meilleurs résultats pour la sélection de l'échelle caractéristique. Ces résultats ont déterminé le choix de la mesure du Laplacien pour notre algorithme de détection de points.

1.2.2 Détecteurs de points d'intérêt

Dans le chapitre 4 nous présentons notre approche pour la détection de point d'intérêt. La première méthode permet d'extraire des points d'intérêt de façon invariante aux changements d'échelle. La deuxième approche est invariante aux changements affines y compris aux changements d'échelle importants. Ces deux détecteurs constituent la contribution principale de la thèse.

Détecteur invariant aux changements d'échelle. Dans la section 4.1 nous décrivons notre approche pour la détection de points d'intérêt invariants aux changements d'échelle. La méthode de la détection est fondée sur deux résultats de la recherche dans le contexte de changements d'échelle : 1) Les points d'intérêt peuvent être adaptés à l'échelle et donnent des résultats répétables [30]. 2) Les extrema locaux des dérivées normalisées dans la dimension d'échelle indiquent la présence de structures locales caractéristiques [70]. Dans un premier temps on détecte des points d'intérêt à plusieurs niveaux d'échelle avec la mesure de Harris [48] qui est basée sur la matrice de deuxième moment. Ainsi, on obtient un ensemble de points d'intérêt qui signalent les endroits où le signal est le plus informatif. Ensuite, on sélectionne les points où la mesure locale (Laplacien) donne une réponse maximale dans la dimension d'échelle. Cela nous permet de rejeter les points les moins discriminants détectés dans le premier stade de l'algorithme. Cette méthode permet de choisir un ensemble de points discriminants pour lesquels les échelles locales sont connues. Les descripteurs calculés sur ces points sont invariants aux changements d'échelle, à la rotation et à la translation.

Détecteur invariant aux changements affines. La section 4.2 présente une méthode de détection de points invariants aux changements affines. L'espace d'échelle construit avec des filtres gaussiens uniformes est souvent utilisée dans le contexte d'extraction de caractéristiques locales invariantes aux changement d'échelle [30, 67, 73, 84] Nous proposons une solution pour la détection de points d'intérêt dans l'espace affine construit avec des filtres gaussiens affines. L'approche est basée sur la méthode introduite pas Lindeberg et Garding [71] qui estime de façon itérative la déformation affine du voisinage d'un point. Le détecteur invariant aux changements affines est une version étendue du détecteur de Harris-Laplace décrit dans la section précédente. La localisation de points est effectuée par le détecteur de Harris basée sur la matrice du deuxième moment. L'échelle caractéristique est indiquée par les maxima locaux de la mesure du Laplacien. Etant donné le point et l'échelle initiale, nous utilisons les propriétés de la matrice de deuxième moment pour déterminer la région affine centrée dans le point d'intérêt. Une procédure itérative est appliquée à chaque point initial fourni par le détecteur Harris multi-échelle. Cette procédure modifie simultanément la position, l'échelle ainsi que le voisinage affine du point. La méthode permet de converger vers un point stable qui est réellement invariant aux changements affines, même si le point initiale est relativement loin de la solution

finale (cf. figure 4.7).

Evaluation comparative de détecteurs. Dans la section 4.3 nous montrons que la performance de nos méthodes est meilleure que celle des approches proposées récemment dans la littérature. Nous avons évalué les détecteurs en utilisant un critère de la répétabilité introduit par [108]. Ce taux mesure le pourcentage de points d'intérêt qui sont répétés entre deux images, par rapport au nombre de points détectés dans la région commune de deux images. Deux points se correspondent si l'erreur de localisation ne dépasse pas $1,5$ *pixel* et si la différence relative entre les surfaces couverts par les voisinages de points correspondants est inférieure à 20%. Le voisinage d'un point est déterminé par l'échelle à laquelle le point est détecté. Pour nos expérimentations, nous avons utilisé plusieurs séquences d'images réelles. Chaque séquence est constituée d'images qui représentent une scène à des résolutions différentes ou prise de points de vue différents. Le facteur d'échelle varie entre 1,2 et 4,5. L'angle de changement de point de vue varie entre 0 et 70 degrés. Les figures 4.9 et 4.11 présentent le taux de répétabilité calculé pour les séquences du test. Dans le cas de changements d'échelle les meilleurs résultats sont obtenus pour la méthode de Harris-Laplace. Ils peuvent s'expliquer par une bonne répétabilité du détecteur de Harris adapté à l'échelle et la bonne performance du Laplacien en sélection d'échelle caractéristique. Dans le contexte de changements perspectifs le détecteur invariant aux changements affines obtient le meilleur score.

1.2.3 Description locale de points d'intérêts

Une étape importante de l'algorithme consiste à capturer l'information portée par les points d'intérêt et à l'utiliser pour mettre en correspondance ces points. Ce problème est étudiée dans le chapitre 5.

Descripteurs différentiels. La section 5.2 décrit la famille de descripteurs basée sur les dérivées locales. Les descripteurs utilisés dans nos algorithmes sont calculés par les filtres orientables [38]. Pour caractériser un point d'intérêt nous utilisons des dérivées de niveaux de gris calculées jusqu'à l'ordre 4. Afin d'obtenir des dérivées indépendantes de la rotation existante entre deux images, la direction de calcul des dérivées est rapportée à la direction dominante dans le voisinage d'un point. Nous avons proposé et évalué une nouvelle méthode pour l'estimation de l'orientation dominante. Une rotation préliminaire du support de calcul rapportée à cette orientation permet d'obtenir l'invariance à la rotation pour un descripteur quelconque. La ressemblance entre les descripteurs est mesurée à l'aide de la distance de Mahalanobis. La matrice de covariance, nécessaire pour calculer la distance, est estimée statistiquement en suivant les points d'intérêt dans des images d'une séquence. Nous avons aussi étudié l'influence de différents types du bruit sur la description locale. Nous avons démontré que les descripteurs différentiels sont particulièrement sensibles au bruit de hautes fréquences et aux erreurs de la localisation de points. L'étude de la matrice de covariance a montré que les invariants différentiels sont corrélés. Pour remédier à ce problème, nous avons proposé une méthode d'apprentissage de la mesure de similarité qui permet d'éliminer certains types de corrélation entre les invariants différentiels.

Entropie de descripteurs. Une de caractéristiques importantes d'un point d'intérêt est la quantité de l'information portée par le voisinage d'un point. Le caractère discriminant du point dépend de cette information et peut être mesuré par l'entropie (cf. 5.3). L'entropie mesure la dispersion des descripteurs de points dans l'espace. Plus la distribution des descripteurs est uniforme plus les descripteurs sont discriminants. Nous avons comparé les points d'intérêt extraits par les détecteurs différents proposés récemment dans la littérature. Les descripteurs différentiels ont été calculés pour chaque ensemble de points extraits par ces détecteurs. Ensuite pour chaque ensemble de descripteurs nous avons calculé l'entropie de la distribution de ces descripteurs. Les meilleurs résultats sont obtenus par les points détectés par notre méthode invariante aux changements d'échelle. Ces points sont plus caractéristiques car ils sont détectés dans l'espace d'échelle à des endroits où le changement du signal est très important dans chaque dimension. Les points invariants aux transformations affines sont moins discriminants car l'information sur la déformation affine a été éliminée. La comparaison a aussi montré que les descripteurs calculés par les filtres orientable [38] sont plus informatifs que les invariants différentiels [60]. Ceci peut s'expliquer par la corrélation des invariants différentiels. Cette étude a permis de choisir le descripteur plus robuste et plus discriminant pour notre algorithme d'appariement.

1.2.4 Appariement et reconnaissance d'images

Le chapitre 6 présente l'algorithme de mise en correspondance et d'indexation, ainsi que les résultats expérimentaux.

Algorithmes d'appariement. La section 6.1 présente l'algorithme de mise en correspondance et de recherche dans une base d'images. La mise en correspondance consiste à trouver des points correspondants entre deux images d'une même scène, mais prise de points de vue très différents. Les points correspondants sont ensuite utilisés pour calculer la transformation géométrique entre les images. Pour effectuer un appariement robuste, dans un premier temps nous déterminons des correspondances de points entre deux images. Pour chaque descripteur de point d'une image nous cherchons le descripteur le plus ressemblant dans l'autre image. Si la distance entre deux descripteurs est supérieure à un certain seuil, les points appariés sont rejetés. Dans le cas où plusieurs descripteurs correspondent à un seul descripteur dans la deuxième image, on garde la paire la plus ressemblante. Ceci permet d'obtenir un premier ensemble de correspondances. Pour pouvoir distinguer les points correctement appariés de ceux qui ne le sont pas, nous ajoutons à posteriori une contrainte globale. Une estimation robuste de la transformation entre deux images, fondée sur la méthode *RANSAC*, permet de rejeter les faux appariements. Dans le contexte de la reconnaissance, la recherche d'images dans une base est effectuée à l'aide de la technique de vote. Chaque point de la base est comparé à la liste des points extraits de l'image requête. Un vote est ajouté à une entrée d'une table de vote si la distance de similarité entre le point de la base et un point de la liste est inférieure à un seuil. On obtient une table de votes où les meilleurs scores correspondent aux images les plus similaires à l'image requête.

Résultats expérimentaux. Nous présentons deux exemples d'applications pour notre

approche. Nous avons effectué un test d'appariement et un test d'indexation. Quelques exemples d'appariement sont présentés dans la section 6.2. L'algorithme invariant aux changements d'échelle permet de mettre en correspondance des images avec un changement d'échelle important. Dans la figure 6.1 les images présentent un changement d'échelle de facteur 4,9. La méthode invariante aux changements affines permet d'aborder les problèmes de changements de point de vue très importants (cf. figure 6.3). Cette méthode est également robuste aux occultations et changements d'éclairage. Dans la section 6.3 nous présentons les résultats d'indexation et de recherche dans une base d'images. La base est constituée de 5000 images. Les images proviennent d'une séquence vidéo de journaux télévisés. Il y a 2 539 342 descripteurs dans la base. Nous avons inclus dans la base une image par séquence de test. Les autres images des séquences ont servi pour évaluer la performance de l'algorithme de reconnaissance. Les figures 6.6 et A.18 montrent les images requêtes pour lesquelles les images correspondantes ont été correctement retrouvées. Les résultats statistiques ont été calculés pour plusieurs séquences de test et sont affichés dans les tableaux 6.1 et 6.2. Nous avons constaté que la méthode Harris-Laplace donne des résultats fiables jusqu'à un facteur d'échelle de 4,4 et la méthode Harris-Affine jusqu'a un changement de point de vue de 70 degrés. Les résultats de la mise en correspondance par rapport aux résultats du test de répétabilité ont montré que les descripteurs différentiels ne permettent pas de trouver tous les points correspondants extraits par les détecteurs. Un descripteur plus robuste et plus discriminant ainsi que des contraintes géométriques permettront d'améliorer les résultats.

1.2.5 Reconnaissance de classes d'objets

Dans le chapitre 7 nous proposons une méthode de détection de visages multiples dans une séquence vidéo.

Détecteur de visages. Dans la section 7.2 nous présentons notre implémentation du détecteur de visages proposé par Schneiderman et Kanade [110]. Le détecteur est basé sur des histogrammes locaux de coefficients calculés par la transformation en ondelettes. Afin de détecter des visages indépendamment du point de vue, nous avons appliqué deux détecteurs. L'apprentissage du premier a été effectué sur les visages vus de face et le deuxième sur les visages vus de profil. L'algorithme calcule la probabilité de l'occurrence de visages dans toutes les positions de chaque image de la séquence. La probabilité est calculée pour plusieurs échelles et pour deux poses; face et profil. Le paramètre de la pose indique si le visage est vu de face, de profil ou d'une pose intermédiaire. Les paramètres qui caractérisent le visage sont alors : la position, l'échelle, et la pose. Chaque détecteur contient aussi un modèle de non-visage qui représente les apparences d'objets différents des visages. La probabilité de présence d'un visage dans une région est obtenue par le calcul de la similarité entre la région et les modèles. La mesure de similarité est basée sur la probabilité accumulée, calculée séparément pour chaque attribut du visage. Ceci permet d'obtenir une robustesse par rapport aux occultations partielles et aux ombres. Toutes les positions dans l'image sont examinées à plusieurs échelles pour détecter les visages de taille différente. La carte de probabilité est alors obtenue pour l'image à des échelles différentes

et les maxima locaux correspondent à des positions de visages. Si la valeur d'un maximum est supérieure à un seuil, on constate une présence du visage. Pour éliminer les collisions entre des occurrences à des échelles différentes nous prenons en compte la taille de la boîte englobante et la probabilité de l'occurrence. La taille de la boîte est associée à l'échelle de détection.

Détection de visages dans une séquence vidéo. Une nouvelle méthode de détection temporelle des visages dans une séquence vidéo a été proposé dans la section 7.3. Les paramètres qui caractérisent le visage peuvent être calculés par un simple détecteur mais la réponse de ce détecteur peut être influencée par des effets différents (occultation, conditions d'éclairage, poses de visage). Sans aucune information supplémentaire ces réponses peuvent être facilement rejetées même si elles sont toujours dues aux présences des visages. Ceci est du au seuil constant de classification. Les exemples (cf. section 7.4.3) montrent que dans le cas d'un seuil trop faible on obtient beaucoup de fausses alarmes. Nous utilisons l'information temporelle, inhérente dans la séquence vidéo, dans la phase de la prédiction et de la mise à jour des paramètres de détection. Une information à priori sur la position et l'échelle fournie par la phase de prédiction accélère la détection. Elle augmente aussi la robustesse de la méthode dans les images ou la probabilité d'occurrence diminue. L'accumulation de la probabilité au cours du temps permet de stabiliser la détection et de la rendre indépendante du seuil de classification. Nous utilisons le filtre de Condensation [52] pour propager les paramètres de détection au cours du temps. Les maxima locaux de la carte de probabilité sont utilisés pour initialiser la procédure. Un maximum local à l'échelle donnée indique une région susceptible de contenir un visage. Notre procédure de détection est divisée en deux phases. La première phase est la détection qui calcule la probabilité de la pose à l'échelle donnée pour chaque position dans l'image. Cette phase est décrite dans la section 7.2. La deuxième phase est la prédiction et la mise à jour de paramètres. Cette partie utilise les probabilités pour suivre les visages dans la séquence. La propagation temporelle est décrite dans la section 7.3.1.

Résultats expérimentaux. Nous avons mené les tests sur des séquences différentes contenant plusieurs visages (cf. section 7.4). Les visages apparaissent à des échelles, à des poses et à des positions différentes. Certains visages disparaissent de la séquence. L'objectif des tests est la comparaison de la détection simple avec la détection temporelle. Dans les séquences les maxima locaux de fausses alarmes disparaissent au cours du temps. Ils peuvent être éliminés car leurs probabilités diminuent rapidement. Deux détecteurs sont appliqués à la séquence de changement de pose. Si la pose change de vue de face en vue de profil la probabilité de profil augmente et la probabilité de face diminue dans des images consécutives. Les deux détecteurs intégrés dans le cadre temporel sont suffisants pour suivre un visage qui change de pose. Le détecteur instantané ne détecte pas de visage si le maximum local de probabilité est trop faible par rapport au seuil de classification. Dans le cas du détecteur temporel ce maximum est choisi pour l'initialisation, ensuite il est propagé et augmente dans les images suivantes. Les faux maxima sont aussi utilisés pour l'initialisation, mais leurs probabilités diminuent à zéro dans les images suivantes, et par conséquent les fausses alarmes sont éliminées. Le calcul de la moyenne de tous les

échantillons de paramètres permet de stabiliser les résultats. En effet, les changements de la position et de la taille des visages sont lisses entre les images consécutives. Ainsi on élimine les discontinuités visibles dans la séquence de détection image par image. Les résultats obtenus par la détection temporelle sont meilleurs. Il n'y a pas de fausses alarmes ni de visages qui ne sont pas détectés, et la détection est stable dans toute la séquence. Toutes ces observations montrent que le détecteur temporel apporte une amélioration importante à la détection des visages dans des séquences vidéo.

1.3 Conclusion et perspectives

La reconnaissance par apparences locales est un domaine relativement nouveau. Un des problèmes les plus difficiles est de rendre la reconnaissance robuste aux transformations géométriques telles que les changements d'échelle importants ou les transformations perspectives. Nous avons apporté des solutions théoriques et génériques à ces problèmes dans le domaine de la détection de caractéristiques locales d'image.

Conclusion. Dans le cadre de ce doctorat, un algorithme d'indexation d'images invariant aux changements d'échelle a été développé. La contribution principale de cet algorithme est la combinaison de deux techniques de détections qui donnent des résultats meilleurs que chacune d'elles appliquée séparément. Les images sont caractérisées par des ensembles de descripteurs calculés en des points caractéristiques détectés automatiquement. Ces descripteurs permettent ensuite d'indexer des images en étant invariant aux rotations, translations et changements d'échelle même importants. Les résultats expérimentaux montrent une excellente performance de la méthode jusqu'à un facteur d'échelle de 4,4 pour une base de 5000 images. Ce travail a été complété par une étude comparative des détecteurs invariants à l'échelle. Les meilleurs résultats ont été obtenus par notre approche.

Les images réelles prises dans des conditions arbitraires sont souvent déformées par une transformation perspective. Sachant qu'on peut approximer localement une transformation perspective par une transformation affine, nous avons proposé un algorithme de détection de points invariants aux transformations affines. Ces points permettent d'apparier des images prises avec un changement de point de vue très important. Les résultats obtenus pour la mise en correspondance ainsi que pour l'indexation d'images sont très prometteurs. Une étude comparative avec les approches existantes a montré que notre approche permet d'obtenir de meilleurs résultats.

Dans la deuxième partie de la thèse, une méthode innovante de détection des visages dans la séquence vidéo a été développée. La détection n'est pas seulement limitée aux vues de face mais permet aussi de détecter les profils de visages. Les attributs caractéristiques des visages sont décrits par les distributions de coefficients obtenus par la transformation en ondelettes. Nous avons rendu cette détection de visages robuste dans des séquences vidéo en utilisant la continuité temporelle des images. Les contributions principales de ce travail sont: 1) l'accumulation des probabilités de détection dans une séquence pour obtenir une détection cohérente au cours du temps, 2) la prédiction des paramètres de position, de l'échelle (taille) et de la pose du visage. 3) la représentation de la pose. Cette représentation est basée sur la combinaison de deux détecteurs, un pour la vue de face et un

pour la vue de profil. Les résultats expérimentaux montrent une amélioration importante par rapport à une détection instantanée.

Perspectives. Les perspectives du travail s'inscrivent dans deux axes de recherche: 1) l'appariement d'images, et 2) la reconnaissance de classes d'objets.
Appariement d'images. La méthode d'appariement que nous utilisons actuellement est basée sur des descripteurs locaux d'images calculés sur des primitives simples contenues dans ces images. La qualité des primitives extraites a une forte influence sur les performances des étapes ultérieures de l'indexation et de la recherche d'image. Des nouveaux types de primitives (régions, contours, coins) et la combinaison de ces primitives peuvent enrichir la description locale d'images. Les performances de l'appariement peuvent être améliorées par l'utilisation combinée de nombreuses sources d'information, notamment celles représentées par la distribution de la couleur et de la texture.

Pour rendre l'appariement plus fiable on peut également appliquer des techniques statistiques. La probabilité d'un appariement correct est alors associée au nombre d'appariements potentiels et à la mesure de similarité entre les points. Ceci peut se faire dans un cadre bayésien.

L'invariance aux changements des conditions d'éclairage reste un problème ouvert. De même, les relations spatiales et les contraintes de voisinage des points sont aussi des sujets importants à étudier.

La construction des modèles à partir de vues multiples permettra de combiner les caractéristiques locales dans un seul modèle d'objet. Le modèle sera basé sur la structure à partir du mouvement et représentera l'espace d'apparence d'un objet. L'espace d'apparence peut être représenté par une base de caractéristiques locales appartenant à un objet. Les apparences d'objet peuvent également être modélisées par une distribution de probabilités.

L'extension du travail vise aussi à valider les approches que nous avons proposées, dans des applications différentes. Il existe un grand nombre d'applications industrielles auxquelles nos approches peuvent apporter une amélioration. Parmi d'autres on peut citer la recherche d'images par le contenu dans les bases d'images, l'appariement d'objets et de séquences dans une vidéo, la détection d'objets dans une vidéo, la navigation des robots, la reconstruction de scènes à partir des images, la construction de mosaïques d'images, des vues panoramiques etc.
Reconnaissance de classes d'objets. Tandis que les approches existantes pour l'appariement donnent des résultats satisfaisants, la reconnaissance de classes d'objet nécessite une étude théorique approfondie. Les études effectuées dans le domaine de la description locale d'image nous permettront de construire des techniques plus performantes pour la description d'objets. Les attributs locaux d'objets non-rigides sont présents à des niveaux d'échelle différents. L'analyse de la représentation multi-résolution permettra de trouver des caractéristiques discriminantes et invariantes aux changements géométriques. Les caractéristiques peuvent être décomposées sur des fréquences de base afin de construire une description compacte.

Les méthodes statistiques peuvent être appliquées pour la sélection de descripteurs pertinents. En particulier, on peut utiliser des techniques comme AdaBoost [100] qui pondèrent les classificateurs pour obtenir une meilleur taux de classification pour la base

d'apprentissage. On peut également utiliser la méthode des Support Vector Machines [20] pour obtenir des partitions optimales de l'espace des caractéristiques.

L'approche hiérarchique et la sélection de caractéristiques plus discriminantes rendront la détection plus fiable et plus rapide. Le processus de décisions sera divisé en plusieurs étapes et les décisions préliminaires seront basées seulement sur des attributs grossiers. L'utilisation de l'information temporelle permettra de rendre plus robuste la reconnaissance dans les séquences vidéo. Une application directe sera la détection de visages ou de personnages dans des séquences vidéo. C'est un sujet relativement nouveau et difficile, mais les premiers résultats sont prometteurs.

De telles approches seront nécessaires pour indexer les contenus multimédia qui sont composés essentiellement d'images et de séquences vidéo.

Chapter 2

Introduction

RECOGNITION is one of the most important problems in the domain of artificial intelligence. The possibilities of human visual perception still far exceed those of artificial vision. The abundance of digital images in the real world requires high-performance and automatic tools to provide fast and reliable navigation in the data. The use of local features in the context of recognition provided for a large progress in terms of the robustness, efficiency and quality of results. The use of interest points for content based object recognition and image retrieval was pioneered by Schmid and Mohr [107], and this has proved to be a very successful and influential approach. In this thesis we focus on the problems of detection of reliable local features, which provide for robust matching and recognition independently of viewing conditions.

In this chapter we present the principal problems, the objectives and the work done during the thesis. At the end of the chapter we present the overview of this manuscript.

2.1 Principal issues

The principal issue of image recognition is to extract semantic context from images. The images in computer vision are represented by pixels. Pixels have different colors, intensities and are ordered in two dimensional matrices. Very short distances between the points trick human visual perception and we see the image as a whole. However, for a computer system, this set of points remains a set of meaningless numbers. The problem is to handle the passage from pixels to semantic content of the image. Global approaches based on color or texture distribution analyze the image as a whole. Unfortunately, they are not robust against occlusions, background clutter and other content changes, which are introduced by arbitrary imaging conditions. An efficient approach which provides for a large progress in solving these problems is based on local features. The local features are local image structures formed by pixels of high intensity variation. Points where the image

content locally changes in two directions are called interest points. These points convey more information due to signal changes, they are therefore more representative for the image. The interest points are then used to represent the content of images. The principal objective of this study is to develop an interest point detector capable of dealing with significant image transformations. The points extracted by this detector are used for matching and recognition. These applications are related but the objectives and the approaches are slightly different. The difference is mainly in the complexity of the approaches.

Our work can be divided into two complementary parts. The first involves matching wide baseline images and recognition of objects, which are deformed by rigid transformations. The approaches for searching the same objects or the scenes, which are viewed in different conditions are in fact different to the approaches for searching a class of objects. The second part of the thesis concerns the recognition of object classes.

2.1.1 Matching and recognition of objects

In general, in this work we follow Schmid and Mohr [107] matching and indexing approach. We first detect a set of interest points that are characteristic for an image. We then compute a description for each interest point. The descriptors are used to find the similar local structures in images. In the case of matching, the descriptors are used to determine point-to-point correspondences. The point-to-point correspondences can be used to recover the geometry between the images. Many applications rely on this geometry. It enables the scene represented by images to be reconstructed or a new view of the scene to be synthesized. A panoramic view can be obtained given a sequence of images of the same scene and the geometric relations between them. Point-to-point correspondences can also be used to localize the camera position and therefore the observer. This opens up numerous opportunities in the context of navigation and visual servoing. The automatic navigation of mobile objects is a wide field of applications as well as the identification and the localization of objects in a scene.

In the case of recognition the point descriptors are applied to find an image model, which is similar to the query image. The models are usually contained in a database. The retrieval of images from databases is one of the principal applications of our approach. Our method can be used to automatically associate the relevant content based description with images. Presently, professional databases still use textual description manually introduced by an operator, which limits the efficiency and the accuracy. There are many different databases, for which the local features can provide reliable retrieval. There are, for example, general databases in photo agencies, press, television, cinemas, private photo collections or specialized databases, such as paintings, trademarks, medical images, product catalogues. The intelligent navigation in these databases is much required.

The robustness and the invariance of the entire matching and recognition process relies on the characteristic points. Therefore in this thesis we focus on the problem of extracting interest points independently of viewing conditions. One of the most frequent image transformations introduced by arbitrary viewing conditions is the perspective transformation. The real images generally represent scenes containing partially smooth objects. The smooth surface can be approximated by a piecewise planar surface. The planar surface

undergoes perspective transformation when viewed from different angles. Finally, the perspective transformation can be locally approximated by affine transformation. Therefore, we assume that local features extracted in an affine invariant way can provide for reliable recognition of locally smooth objects without any constraints on viewing conditions. Without loss of generality, we allow therefore for combined photometric and geometric affine transformations. The geometric transformations include significant scale changes. In the following we discuss the problems to solve.

Interest point. An interest point is represented by a small local neighborhood and is defined by the coordinates in the image, the size and the shape of the structure. These can be affected by different transformations. Rotation, scaling, perspective deformations as well as changes in pixel intensities are the most frequent image transformations. The localization of an interest point is determined by the coordinates relative to the point of origin in the image. An interest point detector should provide the accurate location of points detected in transformed images otherwise the point neighborhoods do not correspond to each other. The scale, namely the size of a point neighborhood is the second important parameter to estimate. We emphasize the scale problems in the next paragraph. Finally, each structure has a specific shape which can be deformed under arbitrary viewing conditions. The problem is to determine the shape of the point independently of these conditions.

Scale invariance. The scale of a local structure is related to the resolution at which the structure is represented in an image. Given a local structure, there exists a minimal resolution, below which the structure is meaningless and a maximal resolution, which depends mainly on the constraints defining the local character of the structure. The problem is to select the appropriate scale at which the structure is most representative. We consider that the effect of changing the camera position along the focal axis or the settings of focal length can be modeled by scaling of local features. Therefore the invariance to scale changes is crucial in our approach. Without the property of scale invariance the complexity of a matching or recognition algorithm is high in the case of significant scale changes. The common region viewed in images being matched represents a small percentage of the coarse resolution image. All the features detected outside of this region are useless. The resolution of the coarse scale image has to be sufficiently high to obtain a stable set of representative points, which means that the size of images being matched has to be large. This also increases the complexity of the algorithm and the time of computation. The terms *scale* and *resolution* are considered equivalent in this manuscript, although, their signification is slightly different. The resolution is determined during the acquisition of images by the parameters of the camera or the scanner, and cannot be artificially increased, although it can be decreased by smoothing and sampling. The scale is, in fact, the factor of relative change in the size of a local structure represented in two images with different resolution. Therefore the term *scale* is always related to the resolution at which the structure is presented. The goal is to select the scale, which is related to the size of the structure.

Affine invariance. An important problem to solve is the influence of an affine transformation on the principal parameters, which define the appearance of a local structure, that is the localization, the size, and the shape. It is essential to determine the affine deformation of the structure before we precisely estimate the other parameters. The translation

of scenes or an object within a scene is less important as the local approach enables the spatial shift of features to be removed. All steps of the detection algorithm should also be invariant to rotation as the camera can be rotated by an arbitrary angle. An affine transformation includes a scale change, rotation and translation. We can handle the affine transformation by solving separately each of the problems, which are related to scale changes, rotation or translation.

Local description. Once the local features are identified in the image, their properties must be captured by descriptors. There are many possibilities to describe the local image structures. The description is necessary for comparing and finding similar structures. The problem is to compute a complete representation that is simultaneously compact and easy to manipulate. The description should be invariant to possible photometric and geometric image transformations. Given the affine covariant regions we can compensate for the affine geometric deformation and compute an affine invariant descriptor. However, the invariance to rotation and illumination changes must also be handled in the process of description. We consider that an affine model of illumination change sufficiently well approximates the real transformation of gray level pixel intensities. Our descriptors are based on local derivatives, and are invariant to affine photometric and geometric transformations of images. We use the responses of several differential measures applied at characteristic points. The set of responses provides a compact and complete local description that is used to compute a similarity between points.

Matching and recognition. A reliable similarity measure is required in every application related to matching or recognition. To find point-to-point correspondences we apply the measure, which is determined by the type of descriptors. The descriptors and the distance measure are necessary but not sufficient to obtain correct point-to-point correspondences. The correctness of the correspondences has to be verified by an additional algorithm that takes into account a global geometric relation between the images. This relation is estimated on the erroneous data, therefore a robust algorithm must be applied. We use the classical RANSAC, which robustly estimates the transformation between matched images and rejects the inconsistent matches. In the case of database retrieval the descriptors are used to find similar quantities in the database. The model that has the highest number of similar descriptors is considered as the most similar to the query image. A voting algorithm is applied for this purpose.

Evaluation. It is necessary to validate the approach on real data to obtain reliable results. The synthetic data significantly facilitates the work but the results are in fact not fully representative as many phenomena cannot be modeled nor predicted. A comparative evaluation on a large set of images can clearly show the advantages and the drawbacks of the method with respect to other existing approaches. The comparative evaluation must rely on objective criteria, which are often difficult to define for general problems.

2.1.2 Recognition of an object class

The second part of the dissertation is devoted to the recognition of an object class, the human face. This work was determined by the industrial project *AGIR - General Architecture for Indexing and Retrieval*. The goal was to develop a reliable face detector.

The Internet opens up new possibilities for searching the same objects or for searching classes of objects in images all over the world. The detection techniques based on accumulated statistical information on local appearance can be used for classification of complex and not necessarily rigid objects. The same classification algorithms can be applied for different objects, like faces, pedestrians, animals, cars, buildings etc. The only difference relies in the image examples, which are used to train the classification algorithm. Face detection is an excellent example of classification. An efficient algorithm for face detection finds applications in many domains of computer vision. It is often used in a preliminary process of face recognition, classification of scenes, structuring of a video sequence, visual surveillance, video conferences etc.

The human face is a relatively distinctive structure, but simultaneously complex in its local and global appearance. The interest point descriptors can represent rigid objects well but are not adapted to problems like face detection. The face can take many different forms, colors, expressions, poses etc. Moreover, the geometric relations among the local structures are not rigid. All these variations have to be incorporated in a face model. The geometric and photometric properties should also be captured by the model. Unfortunately, the complexity of such a description is often prohibitive for fast detection algorithms.

Face models. The existing approaches are usually based on two models, which contain information about the possible appearance of faces and non-faces. Creating the models is one of the most important parts of the algorithm. The representative images must be carefully chosen, otherwise the essential characteristics are not properly captured. There is a large number of such characteristics for the human face. These characteristics must be represented by compact descriptors that preserve the relations between features extracted from different locations, scales and frequencies. The objective of a face detector is to determine the location, the size, and the probability of face appearance. Significantly rotated faces are not frequent in the real video therefore the rotation invariance is not essential. On the other hand faces can appear at different scales, in other words the size of a face can be different. Obviously, a model accumulating many descriptors computed at several scales is less distinctive therefore it is better to apply the multi-scale detection with the model representing faces at one scale only.

Another problem occurs when the view of the face is different to the frontal one. An additional detection parameter can define the pose of the face. The detection of the pose involves either a 3D model or another detector for face profiles. The 3D model requires an initialization and efficient optimization algorithms, which fit the model parameters to the observed 2D view. The profile of the face has a completely different appearance to the frontal view and the characteristic image patterns are different. Therefore we can use one face model trained on frontal and profile views, which is less discriminant or several models for different viewpoints. We will show that two detectors are sufficient, one for frontal views and the other one for profiles.

Face detection in a video sequence. We have adapted the classification method introduced by Schneiderman and Kanade [110], which is based on two models: face and non-face model. The face model represents the statistical distribution of characteristic attributes of the face. The non-face model describes the non-face structures. The multi-resolution wavelet representation enables the local attributes of faces to be extracted. Combined wa-

velet coefficients accumulated in the multi-dimensional histograms provide for a compact and robust description. The classification is most difficult for the objects similar to the face. Therefore, the non-face model has to emphasize the attributes, which are similar but not of the face. Consequently, we learn the non-face model on the false detection results.

The similarity measure is based on the accumulated probabilistic distance computed separately for each face attribute enabling the robustness to partial occlusions and shadows to be obtained. We examine every image location with different window size to extract the faces at different locations and scales. The classification decision is based on the thresholded probability.

A video sequence provides us with redundant information. This information can be used to stabilize the results and eliminate the false detections, which usually appear at a given location only in one frame of the sequence. An algorithm capable of predicting and updating the face parameters can accelerate the procedure by limiting the search area.

The temporal information enables out of plane head rotations to be dealt with. The ambiguity in face appearance, introduced by pose changes, is still a challenging task for any detector or tracker. We do not know whether the appearance probability decreases due to occlusions or changes in the pose. The information provided by detections in previous frames can solve this problem.

2.2 Contributions

In this section we present the principal contributions of the thesis.

Interest point detection. In the previous section we describe the problem of a scale change, which frequently occurs when images are taken from a different distance or with different focal settings. We apply the automatic scale selection for each local feature to handle this problem. One of the contributions of this thesis is the experimental evaluation of the scale selection technique. We compare the ability of several differential expressions to select the characteristic scales related to the local image structure. This allows us to choose the scale selection operator that gives the best results. This method is used to estimate the scale of each interest point.

We propose a detection method, which can be applied to images with significant scale changes including weak perspective deformations. This approach combines two detectors, Harris [48] and Laplacian-of-Gaussians [70], both of which have been previously presented in literature, but separately. The combined Harris-Laplace detector provides better results. The Laplacian operator enables the selection of characteristic scales for points extracted with the Harris detector. Thus, the descriptors are computed on the same point neighborhoods in images of different resolutions, and are therefore invariant to large scale changes.

The principal contributions of the dissertation is an interest point detector invariant to affine geometric and photometric transformations, which include large scale changes. It is an extension of the scale invariant detector. We have introduced an isotropy measure to find the affine transformation of a local isotropic structure. Thus, we can compensate for the affine deformation before computing the description. Very few solutions have previously

been proposed to these problems. Moreover, they handle the affine problem only partially and some of the important parameters are not estimated in an affine invariant way. Our algorithm simultaneously adapts all the affected parameters to obtain the interest points, which covariantly change with viewpoint. These points are used to compute the affine invariant descriptors based on local derivatives.

The rotation of a feature is compensated while computing the description. In this context we have proposed a robust method for estimating the dominant orientation in a local neighborhood of a point. A preliminary rotation of an image patch with respect to the dominant orientation enables the invariance to rotation for any kind of descriptors to be obtained.

One of the contributions of our work is an analysis of differential descriptors and discussion on the usefulness of these descriptors for representing local image structures. We show the sensitivity of these description techniques to different types of noise and the correlation of descriptor components. We propose a method for learning the similarity measure on real data and determining the correlations between the descriptor components.

In order to evaluate and compare different detection methods we have proposed the criteria taking into account the most important parameters, which define a local feature, that is the localization, the scale, the shape, and the information content of the point neighborhood. These criteria are used to evaluate the performance of feature detectors. The experimental results show the excellent performance of our detectors.

We have also evaluated the detectors in the context of matching and recognition. We applied the approach for wide baseline matching and for image retrieval from a database. We use a large number of image samples in order to obtain representative evaluation results.

Face detection in a video sequence. To handle the pose change in a sequence of images we propose to use two face detectors. One uses the face model built with frontal views and the other uses a profile appearance model. These two detectors enable the problems with out of plane head rotations to be dealt with. The combined responses of the detectors are used to predict the actual pose of the face.

We propose a novel approach for detecting faces in a video sequence. We optimize and adapt the approach designed for single image detection to the detection in a video sequence. In order to obtain a stable and coherent detection we use the temporal information represented by consecutive images. The principal parameters defining the face that is the location, the size, and the probability of appearance are propagated along the sequence. The Condensation [52] filter is used to predict and update these parameters. This stabilizes the detection, predicts and reduces the search area as well as eliminating the false detections.

2.3 Overview

In this section we briefly describe the contents of each chapter.

In **chapter 2** we introduce the scale-space theory in the context of local features. We focus on the Gaussian function as numerous studies prove that it is the only function to

generate the scale-space representation. The local features are extracted with operators based on Gaussian derivatives. Therefore we explain how to build an image representation with normalized scale-space derivatives. We emphasize the properties of the second moment matrix and show how to measure the isotropy in a point that is used to estimate the affine deformation of a point. Next, we present the properties of the Hessian matrix, in particular, the ability of the trace of this matrix to select the characteristic scale related to the image structure. Automatic scale selection is one of the most important techniques explored in this study. Therefore, we focus our attention on this approach and we experimentally evaluate the differential expressions usually used in this context.

Chapter 3 describes the main contributions of the thesis. We briefly present the related approaches in the context of scale and affine invariant feature detectors. Next, we present our scale invariant interest point detector and show the advantage of using this approach. We explain in detail each step of the affine invariant feature detector and we analyze the points extracted by this method. To carry out a comparative evaluation of the detectors we propose the criteria measuring the essential properties of the interest points. Finally, we present and analyze the evaluation results.

The description of local characteristic structures is emphasized in **chapter 4**. We present existing descriptors used to represent scale or affine covariant features. We focus on different variants of differential descriptors and identify the reasons for their instability. In this chapter we also present a new method for estimating the dominant orientation of the local image pattern. The approach is evaluated and compared to other existing techniques. We also analyze the similarity measure, which is used to compare the descriptors, and identify the source of correlation of differential invariants. Finally, we present the results of a comparative study of the information content of interest points extracted with different detectors.

In **chapter 5** we present the experimental results for two principal applications of our detectors. At the beginning we outline the matching and the indexing approach. Next, we explain in detail consecutive steps of the algorithms. Finally, we present the matching results for the algorithm based on the scale and the affine invariant detector. We also validate our approach on a large set of image samples in the context of database retrieval.

Chapter 6 is devoted to the problem of detection of an object class such as the human face. We first introduce the wavelet transform as a powerful tool for multi-resolution analysis of complex image structures. We describe a set of descriptors used to represent the appearance of the human face. Next, we present our method for learning the object model. The detection examples follow the outline of the classification algorithm. To improve the detection of faces in a video sequence we propose a temporal approach. Finally, we present the examples of temporal detection in images extracted from a video sequence.

In **chapter 7** we summarize the principal results and discuss the opportunities arising from this work.

Chapter 3

Theory of local features

In this chapter we introduce the theory on scale-space representation. In the early eighties Witkin [131] proposed to consider scale as a continuous parameter and formulated the principal rules of modern scale-space theory relating image structures represented at different scales. Since then, scale-space representation and its properties have been extensively studied and important contributions have been made by Koenderink [59], Lindeberg [67] and Florack [34]. The objective of this chapter is to focus on the properties, which are essential for detecting invariant features. Multi-scale representation of data is crucial for extracting local features. These features exist only in a limited range of scales between the *inner* and *outer* scale, that is the smallest and the largest scale of interest. In order to extract a large number of such features we explore the image representation on a wide range of scales. Numerous studies have shown that the Gaussian kernel is optimal for computing the scale-space representation of an image. Therefore, in section 3.1 we focus on the Gaussian scale-space theory. The scale selection technique is widely used in this thesis for detecting the size of a feature. In section 3.2 we analyze this technique and show the results of a comparative evaluation.

3.1 Multi-scale representation

In this section we describe the multi-resolution image representation based on Gaussian filters. Gaussian derivatives are often used to extract characteristic features. In section 3.1.2 we show how to compute stable derivatives of an image and how to normalize them to obtain derivative responses independent of the image resolution. The local features are defined by the location, the scale and the shape, which undergo affine transformation when viewed from different angles. To estimate the affine deformation of an interest point we explore the properties of the second moment matrix, which are described in section 3.1.3. The components of the Hessian matrix can be used to detect a characteristic scale of

a local structure and to describe a shape of the structure, therefore in section 3.1.4 we present this matrix.

3.1.1 Gaussian scale-space

In the discrete domain of digital images the scale parameter is also discretized. Thus, the scale-space representation is a set of images represented at different discrete levels of resolution [131]. Koenderink [59] showed that the scale-space must satisfy the diffusion equation for which the solution is a convolution with the Gaussian kernel. Furthermore he showed that this kernel is unique for generating a scale-space representation. The uniqueness of the Gaussian kernel was confirmed with different formulations by Babaud [5], Lindeberg [65] and Florack [35]. These results lead to the conclusion that the convolution with the Gaussian kernel is the best solution to the problem of constructing a multi-scale representation. The bi-dimensional Gaussian function is defined by:

$$g(\sigma) = \frac{1}{2\pi\sigma^2} \exp^{-\frac{x^2+y^2}{2\sigma^2}},$$

The uniqueness of the family of Gaussian kernels is also determined by the following properties: linearity, separability, causality and semi group property. Extensive discussion on these properties can be found in the literature [35, 59, 68, 115, 131]. The separability enables a multi-dimensional Gaussian kernel to be obtained as the product of one-dimensional kernels:

$$g(x, y) = g(x)g(y)$$

This property is very useful in practice as the smoothing of a two-dimensional signal can be replaced by two one-dimensional smoothings, one for each dimension. A one dimensional Gaussian filter can be implemented as a recursive filter [26], which significantly accelerates the computation process in the case of larger Gaussian kernels (i.e. $> \sqrt{2}$). The causality condition states that no additional, artificial structure should be created when computing the coarse scale image, that is the image at a coarser scale is a simplified representation of the image at a finer scale. The commutative semi-group property states that n successive smoothings of an image give the same result as one smoothing with the kernel size equal to the sum of all the n kernels. Additionally, the n operations can be done in any order:

$$g(\sigma_1) * \ldots * g(\sigma_n) * I(\mathbf{x}) = g(\sigma_1 + \ldots + \sigma_n) * I(\mathbf{x})$$

Usually, a uniform scale-space is used, but a general scale-space representation is computed with affine filters. In the following we explain in detail each of these two representations.
Uniform scale-space. Different levels of the scale-space representation are, in general, created by convolution with the Gaussian kernel (cf. figure 3.1):

$$L(\mathbf{x}, \sigma) = g(\sigma) * I(\mathbf{x})$$

with I the image and $\mathbf{x} = (x, y)$ the point location. The kernel is circularly symmetric and parameterized by one scale factor σ. The semi-group property facilitates the scale-space representation in several ways, which we describe in the following. A coarse scale image

FIG. 3.1: *Uniform Gaussian kernel.*

is obtained by smoothing the fine scale image. This operation is repeated on consecutive coarser levels to obtain the multi-scale representation. In order to accelerate the operation one can sample the coarser scale image by the corresponding scale factor after every smoothing operation. One should be careful choosing the scale and the sampling factor as it may lead to aliasing problems. Moreover, additional relations have to be introduced in order to find the corresponding point locations at different scale levels. This makes any theoretical analysis more complicated. The pyramid representation and its different aspects have been extensively studied by Burt [16], Crowley [21, 22], and Meer [82]. This representation was also used by Crowley [22, 23] and Lowe [73] for local feature detection. A sketch of a scale-space pyramid is presented in figure 3.2(a).

A scale-space representation can also be built by successive smoothing of the high resolution image with kernels of different scales. Building the scale-space is more time consuming, when the scale levels are not sampled. The information is then redundant, but there is no need for computing the corresponding point locations at different scales (cf. figure 3.2(b)). If we keep all points at every scale-space level, we preserve a direct relation between the theoretical analysis and real computations.

The feature detectors are mainly based on simple structure tensors as for example the second moment matrix in the interest point detector, which was introduced by Harris and Stephens [48]. The *inner* and the *outer* scale of interest, within which the structure can be analyzed, are imposed by the acquisition conditions, i.e. the resolution and the field of view. The minimal size of an image structure that can be considered as a characteristic is limited by the resolution and the noise. The *inner* scale is a factor relative to the structure size, in other words is the minimal size of the point neighborhood that represents the essential information about the structure. The *outer* scale, that is the largest size of the structure, is limited by the constraints defining the local character of the feature. It is also limited by the size of an image. The scale factor must be distributed exponentially between the inner and outer limits $\sigma_n = \sigma_0 s^n$, in order to maintain a uniform change of information between successive levels of resolution.

Affine scale-space. A more general representation is the affine scale-space. The theory of affine Gaussian scale-space is very useful, when dealing with affine transformations of the image patch. This representation can be generated by convolution with affine Gaussian kernels (cf. figure 3.3, equation 3.1).

$$g(\Sigma) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp^{-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2}} \tag{3.1}$$

If the matrix Σ is equal to an identity matrix multiplied by a scalar, this function corresponds to a uniform Gaussian kernel. We deal with a three-dimensional space (x, y, σ)

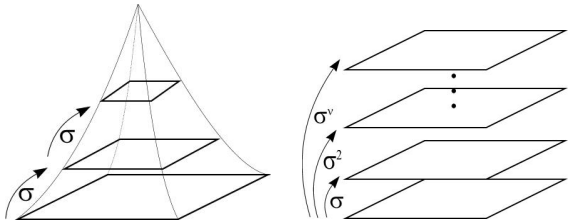


FIG. 3.2: *Scale space: a) Pyramid representation constructed by combined smoothing and sampling. b) Scale-space representation constructed by successive smoothing of the high resolution image.*

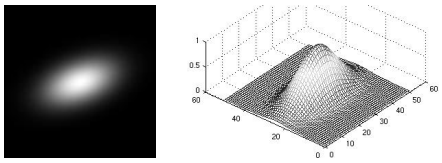


FIG. 3.3: *Affine Gaussian kernel.*

if traditional, uniform Gaussian filters are used. The Gaussian kernel is then determined by one scale parameter σ. If Σ is a symmetric positive 2x2 matrix, the number of degrees of freedom leads to a high dimensional space too complex to deal with. However, the complexity is reduced if we compute the representation for one image point. The affine filters applied in a point were used in practice by Lindeberg [71] for the shape-from-texture problem.

There are several approaches to computing the convolution with an affine kernel. One of the possibilities is to use the Fourier transform of the discrete Gaussian kernel and to perform the multiplication with an image in the frequency domain. Recursive implementation of such filters is proposed in [42]. However, if we compute the filter response in one image point there is no benefit in using the recursive filters. The same computational cost is if we sample the affine kernel and directly convolve it with the image. The method used in the approach described in this manuscript is based on a decomposition of the kernel matrix into a product of rotation and scaling matrices:

$$\Sigma = R^T \cdot D \cdot R = \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix} \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix} \begin{bmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{bmatrix}$$

This equation shows that the affine smoothing corresponds to the convolution with a rotated elliptical Gaussian kernel. To simplify the computation process we can transform the point:

$$\mathbf{x}^T \Sigma^{-1} \mathbf{x} = \mathbf{x}^T \frac{1}{\sigma_x \sigma_y} \Sigma_N^{-1} \mathbf{x} = \mathbf{x}^T \Sigma_N^{-\frac{1}{2}} \frac{1}{\sigma_x \sigma_y} \Sigma_N^{-\frac{1}{2}} \mathbf{x} = (\Sigma_N^{-\frac{1}{2}} \mathbf{x})^T \frac{1}{\sigma_x \sigma_y} \Sigma_N^{-\frac{1}{2}} \mathbf{x}$$

Thus, the affine smoothing is done by the convolution of the uniform kernel $\sigma_N = \sqrt{\sigma_x \sigma_y}$ with the point transformed by $\mathbf{x}' = \Sigma_N^{-1/2} \mathbf{x}$. A problem occurs when σ_x is very different

from σ_y. In this case, the image sampling by intervals σ_x/σ_N and σ_y/σ_N involves some loss of information and introduces artifacts. To overcome this problem we can set the size of the uniform kernel to $\sigma_N = max(\sigma_x, \sigma_y)$ and then compute the matrix $\Sigma_N^{-1/2}$. We then assure that the image is correctly sampled, that is the image is sampled at least once between every two neighboring pixels (cf. figure 3.4).

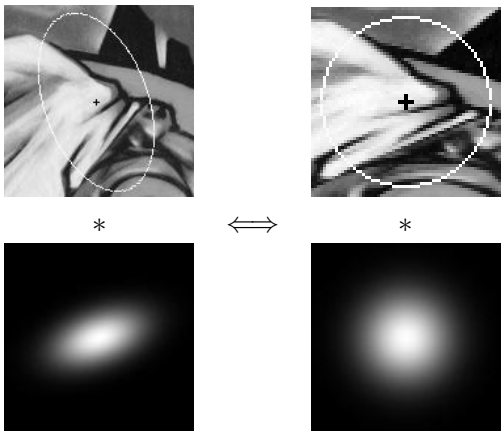


FIG. 3.4: *(a) Smoothing with an affine Gaussian kernel, (b) Equivalent operation with a uniform kernel and the affine transformed image.*

Scale-space representation is not explored directly, but differential operators are often applied to extract the appropriate information. These differential operations are mainly based on Gaussian scale-space derivatives.

3.1.2 Scale-space derivatives

The usefulness of derivatives in the local signal analysis can be illustrated by Taylor expansion. Taylor expansion plays a crucial role in filter design [89] and, up to a chosen order, locally approximates the structure of an image. In general, the expansion is computed up to second order.

$$I(\mathbf{x}_0 + \Delta\mathbf{x}) \approx I(\mathbf{x}_0) + \Delta\mathbf{x}^T \nabla I(\mathbf{x}_0) + \Delta\mathbf{x}^T \mathcal{H}(\mathbf{x}_0) \Delta\mathbf{x}$$

We make a rather general assumption that an image is a differentiable two dimensional signal. There is, of course the notorious boundary problem but it can be ignored if the analyzed local structure is sufficiently far from the boundaries with respect to the scale of operators. The components of the second order Taylor expansion, both the gradient and the Hessian matrix are often used separately (cf. sections 3.1.3 and 3.1.4). The properties of the second moment matrix (cf. equation 3.2) have been explored in the context of feature detection in [7, 19, 48, 71, 92]. The usefulness of the Hessian matrix in this domain is less known, although in recent years the components of this matrix appeared to be particularly useful for automatic scale selection [2, 17, 71]. Both matrices are used in this

thesis for detecting local features, we therefore describe in this section how to compute their components.

Gaussian derivatives. A feature (i.e. blob, corner, edge) can be extracted at different resolutions by applying an appropriate function at different scales. The detection functions are mostly based on Gaussian scale-space derivatives, as the linear derivatives of Gaussians are suitable for modeling the human visual front-end. The weighted difference, computed by convolving the original signal with a derivative of the Gaussian, may be seen as a generalization of a difference operator. When the scale parameter tends to zero, the scale-space derivative approaches the true derivative of the function. The aim of the scale-space analysis is to explore an image representation on a wide range of scales in order to extract the salient information.

Uniform derivatives. In general, the filters derived from the uniform Gaussian kernel are used in practice in the image processing domain. The derivatives at different scales can be computed by smoothing the image with the Gaussian and differentiating the smoothed signal. All the properties of the Gaussian kernel also holds for its derivatives. Therefore, if we apply these operations in the inverse order we obtain the same results. Another option is to convolve the image with a derivative of a scaled Gaussian kernel. All these methods are equivalent. Given any image function $I(\mathbf{x})$ the first derivative can be defined by:

$$L_x(\mathbf{x};\Sigma) = \frac{\partial}{\partial x} * g(\Sigma) * I(\mathbf{x})$$

The general expression for Gaussian derivatives is the following:

$$g_{i_1 \ldots i_m}(\mathbf{x},\Sigma) = \frac{\partial^m}{\partial i_1 \ldots \partial i_m} \frac{1}{2\pi\sqrt{\det \Sigma}} \exp^{-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2}}$$

where m is the derivative order and i the Cartesian coordinate in the image. In the case when Σ is an identity matrix multiplied by a scalar we deal with traditional uniform Gaussian derivatives. Figure 3.5 shows the derivative kernels up to the fourth order. The complementary derivatives in the orthogonal direction can be obtained by a simple rotation of 90 degree.

Normalized derivatives. The amplitude of spatial derivatives, in general, decreases with scale, due to the the response being more smoother on a larger scale. In the case of the structures present at a large range of scales, like a corner or a step-edge, we would hope to have the derivative constant over scale. In order to maintain the property of *scale invariance* the derivative function must be normalized with respect to the scale of derivation. The scale invariance properties are described in [67, 70]. The scale normalized derivative D of order m is defined by:

$$D_{i_1 \ldots i_m}(\mathbf{x},\sigma) = \sigma^m L_{i_1 \ldots i_m}(\mathbf{x},\sigma) = \sigma^m g_{i_1 \ldots i_m}(\sigma) * I(\mathbf{x})$$

In the following we show the necessity of using the normalization factor σ^m. Consider two images I and I' imaged at different scales. The relation between the two images is then defined by: $I(\mathbf{x}) = I'(\mathbf{x}')$, where $\mathbf{x}' = s\mathbf{x}$. Notice that a possible shift of a point is ignored, as it is removed by differentiation. Image derivatives are then:

$$I_{i_1 \ldots i_m}(\mathbf{x}') = s^m I_{i_1 \ldots i_m}(s\mathbf{x})$$

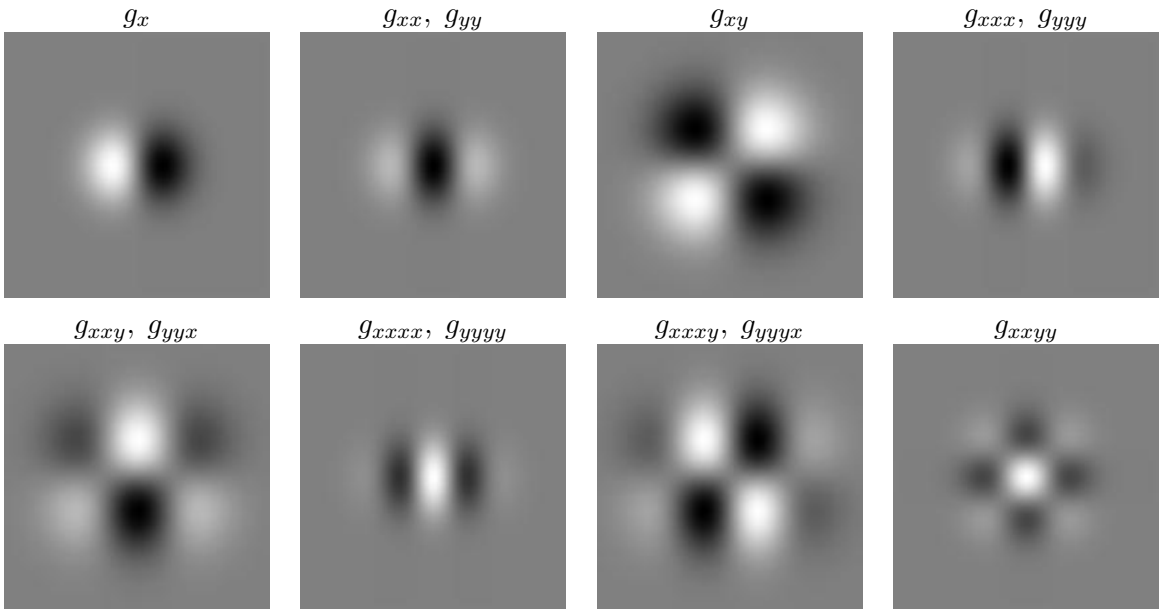


FIG. 3.5: *Uniform Gaussian derivatives up to fourth order. The derivative kernel g_{yy} is equal to the g_{xx} kernel rotated by 90 degree. Similarly, we can obtain g_{yyy}, g_{yyx}, g_{yyyy} and g_{yyyx} kernels.*

If we suppose that the derivative kernel of scale σ is normalized by the same scale factor, we obtain:

$$\sigma^m g_{i_1 \dots i_m}(\sigma) * I(\mathbf{x}) = s^m \sigma^m g_{i_1 \dots i_m}(s\sigma) * I(\mathbf{x}')$$

Thus, for normalized derivatives the response has the same value:

$$D_{i_1 \dots i_m}(\mathbf{x}, \sigma) = D'_{i_1 \dots i_m}(\mathbf{x}, s\sigma)$$

We can see that if we multiply the derivatives by the kernel size we obtain the same derivative values for local structures represented at corresponding scales.

Affine derivatives. The affine derivatives are very useful if we deal with affine invariance. In order to obtain the affine invariance of derivatives of an arbitrary local structure the Gaussian kernels must be adapted to the shape and to the scale of the structure. Thus an anisotropic shape of the feature is not biased by an isotropic smoothing with a uniform filter. To adapt the filter, without any prior knowledge of the shape of the structure, we have to explore many possible combinations of kernel parameters. However, the three degrees of freedom of an affine Gaussian kernel make it difficult to investigate all possible combinations. Therefore, in practice, we constrain the possible shapes of kernels. The computation of affine directional derivatives for all image points can be accelerated with recursive implementation [42]. We apply the non-uniform derivatives only for points of interest, where the x, y coordinates are fixed, there is therefore no need to use the recursive filters. The derivatives are computed by convolving the affine normalized image with sampled deriva-

tive of uniform Gaussian kernel (cf. section 3.1, figure 3.4). In the next section we apply these derivatives to compute the components of the second moment matrix.

3.1.3 Second moment matrix

The second moment matrix described in this section is often used for feature detection or description of local image structures. This matrix is also called the auto-correlation matrix and is defined by:

$$\mu(\mathbf{x}, \sigma_I, \sigma_D) = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} = \sigma_D^2 \, g(\sigma_I) * \begin{bmatrix} L_x^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \quad (3.2)$$

It describes the gradient distribution in a local neighborhood of a point. The gradient derivatives are determined by a local scale σ_D (derivation scale). The derivatives are averaged in the neighborhood of the point by smoothing with a Gaussian window of size σ_I (integration scale). The eigenvalues of this matrix represent two principal curvatures of a point. This property enables the extraction of points, for which both curvatures are significant, that is the signal change is significant in the orthogonal directions. These points are more stable in arbitrary lighting conditions and more representative than other image points. One of the most reliable interest point detectors known as Harris detector [48] (cf. equation 3.21) is based on this principle.

In the affine scale-space, the second moment matrix μ in a given point $\mathbf{x}$ is defined by:

$$\mu(\mathbf{x}, \Sigma_I, \Sigma_D) = \det(\Sigma_D) \, g(\Sigma_I) * ((\nabla L)(\mathbf{x}, \Sigma_D)(\nabla L)(\mathbf{x}, \Sigma_D)^T)$$

where Σ_I and Σ_D are the covariance matrices which determine the integration and the derivation Gaussian kernel. It is impossible in practice to compute the matrix for all the possible combinations of kernel parameters. To limit the number of degrees of freedom we impose the condition $\Sigma_I = s\Sigma_D$, where s is a scalar. Hence, the derivation and the integration kernel will differ only in size and not in shape. It means that the scale factor between two orthogonal directions is the same for smoothing and integrating the derivatives of the second moment matrix.

Affine transformation of a point. The second moment matrix has a property which makes it particularly useful for estimating an anisotropic shape of a local image structure. This property was explored by Lindeberg [67, 71] and later on by Baumberg [7] to find the affine deformation of an isotropic structure. In the following we show how to determine the anisotropic shape. Consider a point $\mathbf{x}_L$ transformed by a linear transformation $\mathbf{x}_R = A\mathbf{x}_L$. The matrix μ_L computed in the point $\mathbf{x}_L$ is then transformed in the following way:

$$\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) = A^T \mu(A\mathbf{x}_L, A\Sigma_{I,L}A^T, A\Sigma_{D,L}A^T)A = A^T \mu(\mathbf{x}_R, \Sigma_{I,R}, \Sigma_{D,R})A \quad (3.3)$$

If we denote the corresponding matrices by:

$$\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) = M_L \qquad \mu(\mathbf{x}_R, \Sigma_{I,R}, \Sigma_{D,R}) = M_R$$

these matrices are then related by:

$$M_L = A^T M_R A \quad M_R = A^{-T} M_L A^{-1} \quad (3.4)$$

The derivation and the integration kernels are in this case transformed by:

$$\Sigma_R = A\Sigma_L A^T$$

Let us suppose that the matrix M_L is computed in such a way that :

$$\Sigma_{I,L} = \sigma_I M_L^{-1} \quad \Sigma_{D,L} = \sigma_D M_L^{-1} \tag{3.5}$$

where the scalars σ_I and σ_D are the integration and derivation scales respectively. We can then derive the following relation:

$$\begin{aligned} \Sigma_{I,R} &= A\Sigma_{I,L}A^T = \sigma_I(AM_L^{-1}A^T) = \sigma_I(A^{-T}M_L A^{-1})^{-1} = \sigma_I M_R^{-1} \\ \Sigma_{D,R} &= A\Sigma_{D,L}A^T = \sigma_D(AM_L^{-1}A^T) = \sigma_D(A^{-T}M_L A^{-1})^{-1} = \sigma_D M_R^{-1} \end{aligned} \tag{3.6}$$

We can see that imposing the conditions 3.5 entails the relations 3.6 under the assumption that the points are related by an affine transformation. We can now inverse the problem and suppose that we have two points related by an unknown affine transformation. If we estimate the matrices Σ_R and Σ_L such that the matrices verify conditions 3.5 and 3.6, then relation 3.4 is true. The presented property enables the transformation parameters to be expressed directly by the matrix components. The affine transformation can then be defined by :

$$A = M_R^{-1/2} R M_L^{1/2}$$

where R represents an arbitrary rotation. In section 4.2.2 we present an iterative algorithm for estimating the matrices Σ_R and Σ_L. In section 5.2.2 we show how to recover the rotation R in a robust manner. Thus, we can estimate the affine transformation between two corresponding points without any prior knowledge about this transformation. Furthermore, the matrices M_L and M_R, computed under conditions 3.5 and 3.6, determine corresponding regions defined by $\mathbf{x}^T M \mathbf{x} = 1$. If the neighborhoods of points $\mathbf{x}_R$ and $\mathbf{x}_L$ are normalized by transformations $\mathbf{x}'_R = M_R^{1/2}\mathbf{x}_R$ and $\mathbf{x}'_L = M_L^{1/2}\mathbf{x}_L$, respectively, the normalized regions are related by a simple rotation $\mathbf{x}'_L = R\mathbf{x}'_R$.

$$\mathbf{x}_R = A\mathbf{x}_L = M_R^{-1/2} R M_L^{1/2}\mathbf{x}_L, \qquad M_R^{1/2}\mathbf{x}_R = R M_L^{1/2}\mathbf{x}_L \tag{3.7}$$

The matrices M'_L and M'_R in the normalized frames are equal to a pure rotation matrix. In other words, the intensity patterns in the normalized frames are isotropic. This operation is illustrated in figure 3.6. A similar compensation by the square root of the second moment matrix was used in [40].

Isotropy measure. In the following we interpret the second moment matrix, presented above, in terms of the isotropy measure. Without loss of generality we suppose that a local anisotropic structure is an affine transformed isotropic structure. This provides a solution for the problem of affine deformation of local patterns when viewed from different angles. An isotropic structure deformed by the affine transformation becomes anisotropic. To compensate for the affine deformation, we have to find the transformation that brings the anisotropic pattern to the isotropic one. Notice that the rotation preserves the isotropy or the anisotropy of an image patch. Therefore, an affine deformation of an isotropic

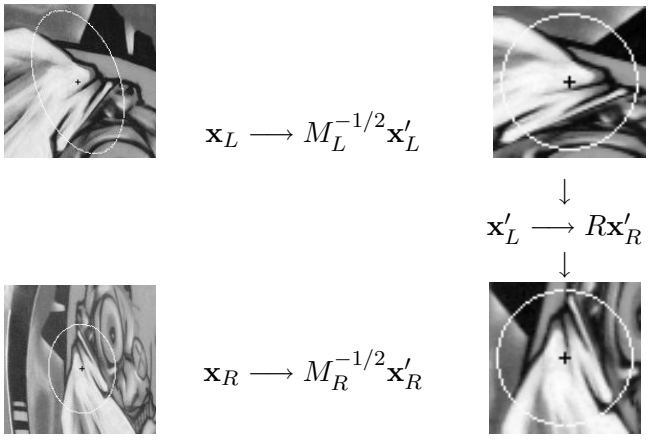


FIG. 3.6: *Diagram illustrating the affine normalization using the second moment matrices. Image coordinates are transformed with matrices* $M_L^{-1/2}$ *and* $M_R^{-1/2}$ *(cf. equation 3.7).*

structure can be determined up to the rotation factor. This factor can be recovered by other methods (cf. section 5.2.2). The second moment matrix $\mu(\mathbf{x}, \sigma_I, \sigma_D)$ (cf. equation 3.2) can be interpreted as the isotropy measure applied in a point $\mathbf{x}$ within a local neighborhood of size σ_I. The local isotropy can be measured by the eigenvalues of the matrix μ. If the eigenvalues are equal we consider the point isotropic. To obtain a normalized measure we use the eigenvalue ratio:

$$\mathcal{Q} = \frac{\lambda_{min}(\mu)}{\lambda_{max}(\mu)} \tag{3.8}$$

The values of $\mathcal{Q}$ vary in the range of $[0 \dots 1]$ with 1 for a perfect isotropic structure. This measure can give a slightly different response for different scales as the matrix μ is determined by two scale parameters. These scales should be selected independently of the image resolution. The scale selection technique described in section 3.2, gives the possibility to determine the integration scale related to a local image structure. The derivation and integration scales can be related by $\sigma_D = s\sigma_I$, where s is a constant factor. For obvious reasons the derivation scale should always be smaller than the integration scale. The factor s should not be too small, otherwise the smoothing is too significant with respect to the derivation. On the other hand s should be small enough, so that σ_I can average the covariance matrix $\mu(\mathbf{x}, \sigma_D, \sigma_I)$ in the local neighborhood. The idea is to suppress the noise without suppressing the anisotropic shape of the observed image structure.

More sophisticated approach is to select the derivation scale σ_D independently of the scale σ_I. Given the integration scale we can search for the scale σ_D, for which the response of the isotropy measure attains a local maximum. Thus, the shape selected for the observed structure is less anisotropic. A similar approach for selecting local scale was introduced by Lindeberg [2, 71], but he proposed selecting the scale for which a normalized anisotropy (cf. equation 3.9) assumed a maximum over scale.

$$\mathcal{Q}_A = \frac{\sqrt{\mathrm{trace}^2\mu - 4\det\mu}}{\mathrm{trace}\mu} \tag{3.9}$$

This measure can be also expressed by the eigenvalues:

$$\mathcal{Q}_A = |\frac{\lambda_{max}(\mu) - \lambda_{min}(\mu)}{\lambda_{max}(\mu) + \lambda_{min}(\mu)}|$$

We notice the similarity between $\mathcal{Q}$ and $\mathcal{Q}_A$. Although, $\mathcal{Q}_A$ tends to zero if the point becomes more isotropic. Contrary to our approach, the image pattern is not affine normalized in the iterative procedure estimating the anisotropic shape (cf. section 4.2.2). Furthermore, in our experiments we noticed that this procedure diverges more frequently if the local scale is selected by a maximum of $\mathcal{Q}_A$ measure.

3.1.4 Hessian matrix

The second 2x2 matrix issued from the Taylor expansion is the Hessian matrix (cf. equation 3.10). This matrix can also be used to describe the properties of local image structures. The Hessian of an image is built with second order partial derivatives.

$$\mathcal{H}(\mathbf{x}, \sigma_D) = \sigma_D^2 \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma_D) & L_{xy}(\mathbf{x}, \sigma_D) \\ L_{xy}(\mathbf{x}, \sigma_D) & L_{yy}(\mathbf{x}, \sigma_D) \end{bmatrix} \tag{3.10}$$

These derivatives encode the shape information by providing the description of how the normal to an isosurface changes. Particularly interesting are the filters based on the determinant and the trace of this matrix. Ter Haar Romeny et al. [120] use the determinant of the Hessian matrix to compute the product of principal curvatures of the intensity surface in an image. The components of this matrix have also been used for extracting local features by Beaudet [8], Kitchen and Rosenfeld [58], and for describing interest points by Koenderink [59], ter Haar Romeny [120] and Schmid [107]. The trace of this matrix denotes the Laplacian filter, which is often used for isotropic edge detection [123]. These and other entities derived from eigenvalues of the Hessian matrix have been used for scale selection in [70]. The importance of the Laplacian of Gaussian (LoG) for bioperception has been emphasized in the work of Marr [80]. In our approach we apply the LoG operator for automatic scale selection of local image structures. This technique is explained and evaluated in the next section. The difference-of-Gaussian (DoG) which is an approximation of the LoG has been successfully applied to feature detection invariant to scale changes [22, 66, 73]. In this thesis we always use the LoG filter as it is the stable implementation of the Laplacian operator. The uniform LoG kernel is presented in figure 3.7. The size of the kernel is parameterized by a scale factor. A general expression for the Hessian matrix in the context of affine scale-space is defined by:

$$\mathcal{H}(\mathbf{x}, \Sigma_D) = ((\nabla\nabla^T L)(\mathbf{x}, \Sigma_D))$$

In practice, the affine derivatives can be computed as explained in section 3.1.

3.2 Automatic scale selection

Scale invariance is one of the objectives in this work, therefore in this section we focus on the methods for determining a scale of a local image structure. Automatic scale selection

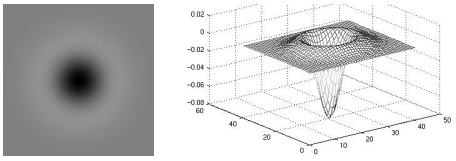


FIG. 3.7: *Laplacian kernel* $g_{xx}(\sigma_D) + g_{yy}(\sigma_D)$.

and the properties of the selected scales have been extensively studied by Lindeberg [70]. The idea is to select the *characteristic* scale, for which a given function attains an extremum over scales. The name *characteristic* is somewhat arbitrary as a local structure can exist at a range of scales and within this range there is no preferred scale of perception. However, for a particular descriptor a scale can be named *characteristic*, if the descriptor computed at this scale conveys more information comparing to descriptors at other scales (cf. section 5.3).

In section 3.2.1 we present the theoretical analysis of the scale selection technique. The influence of normalization parameters on the selected scale is explained in section 3.2.2. Differential expressions often used for feature detection are introduced in section 3.2.3. Finally, section 3.2.4 presents evaluation results of the differential expressions applied to scale selection.

3.2.1 Scale-space maxima

In the following we present the analytical relations between a scale space maximum and the scale of a local image structure. Let F be a function normalized with respect to scale, that we use to build the scale-space. The set of responses for a point $\mathbf{x}$ is then defined by $F(\mathbf{x}, \sigma_n)$ with $\sigma_n = \sigma_0 s^n$. This set is called a scale trace. The factor σ_0 is the initial scale at the finest level of resolution and σ_n denotes successive levels of the scale-space representation. Parameter s enables the exponential increase of scale to be obtained for uniform information change between the scale levels.

The top rows of the frames in figure 3.8 show theoretical signal configurations, for which the Laplacian and the gradient is computed. The images present differentiation kernels in the neighborhood of a corner or an edge. In the middle row we display the scale trace for the corresponding signal configuration. Note that for some of them the gradient does not attain an extremum. This occurs for points near corners and edges if there is no other signal change in the neighborhood. Note that the extremum of the gradient is also less distinctive than the one obtained for the Laplacian. We can expect that in real images the Laplacian attains the extremum over scale more frequently than the gradient. This is confirmed by the experimental results for real images, which are presented in figure 3.9.

We can show for some of the signal configurations displayed in figure 3.8 that the maximum of the normalized derivative is related to the distance from the signal change. The necessary condition to find a local extremum over scale is $\frac{\partial}{\partial \sigma} F_{norm} = 0$. Given a

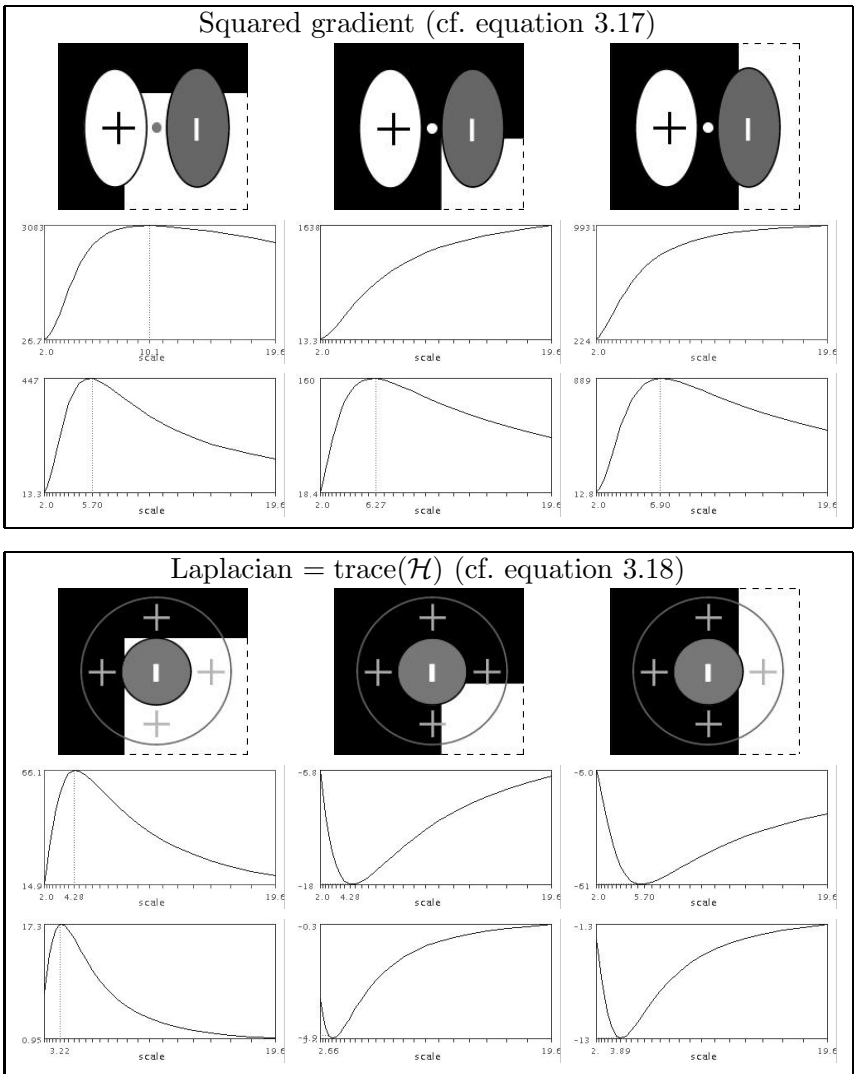


FIG. 3.8: *Scale trace of squared gradient and Laplacian applied to corner and edge models. First column: point inside a corner. Second column: point outside a corner. Third column: point nearby an edge. In the frames: Top rows: Corner and edge models. Middle rows: Scale trace for* $\gamma = 1$. *Bottom rows: Scale trace for* $\gamma = 0.5$.

function representing the step-edge displayed in figure 3.8(c),(f):

$$f_{step-edge}(x) = \begin{cases} 0 & x < x_0 \ \forall y \\ 1 & x \geq x_0 \ \forall y \end{cases}$$

we can compute the convolution of the corner with normalized Laplacian centered in point $(0, 0)$:

$$f_{xx_{norm}}(x, \sigma) = \frac{x}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

The extremum of the function can be found as follows:

$$\frac{\partial}{\partial\sigma} f_{xx_{norm}}(x, \sigma) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} (-\frac{x}{\sigma^2} + \frac{x^3}{\sigma^4}) = 0 \Leftrightarrow \sigma = |x_0| \tag{3.11}$$

For a given point (x_0, y_0) the normalized Laplacian attains an extremum at $\sigma_{extremum} = |x_0|$. The problem can occur when the point is equal $x_0 = 0$, that is at the maximum of the first derivative. Fortunately, the interest points detected by the operators usually applied for this purpose, are not localized in $x_0 = 0$, that is exactly at the corner junction. This is due to the detection scale, which in practice cannot be equal 0, and as we will show in section 4.1.3 the interest points change the location with respect to the detection scale. The detailed mathematical operations can be found in annex A.1. Such responses to edges and corners were also analyzed in [67]. Note that the linear scaling of coordinates x, y involves the same linear scaling of $\sigma_{extremum}$. We can also show that there is no extremum over scale for the first order derivative of a step-edge (cf. figure 3.8(c)).

$$\frac{\partial}{\partial\sigma} f_{x_{norm}}(x, \sigma) = \frac{x^2}{\sigma^3\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \neq 0 \ \forall\sigma \tag{3.12}$$

Thus, we can recover the scale of a feature by looking at the maximum of normalized second order derivatives.

3.2.2 Gamma normalization

The normalization factor, which is applied for computing scale-space derivatives, has a property, which can be particularly useful for parameterizing the mechanism of scale selection. In this section we analyze the influence of the normalization factor on the local maxima over scales. Instead of normalizing the derivatives by factor σ^n, where n is the derivative order, we can apply $\sigma^{\gamma n}$. When $\gamma = 1$ we deal with so-called perfect scale invariance, that is the amplitude of the normalized derivatives is independent of the signal resolution. The invariance is not necessarily preserved in the case of scale selection operators based on combinations of derivatives of different orders. If such operators are used the magnitude of the scale trace can be different for a point represented at different resolutions. The normalization factor should be then $\gamma \neq 1$ to preserve the magnitude invariance. Nevertheless, the local maxima over scales are still preserved, even for $\gamma \neq 1$. Given a γ-normalized detection operator, there is a certain range of γ values within which a specific image structure has assigned a characteristic scale. The extremum of derivatives

over scale has a tendency to move to lower scales with decreasing γ factor. This effect is illustrated in middle and bottom rows of figures 3.8 and 3.9. By setting $\gamma < 1$ we can obtain an extremum for a lower σ value. Therefore, we can explore a narrower range of scales when searching for the local maximum. Lindeberg [71] showed that for a simple, periodical, one-dimensional signal, like sine or cosine, the maximum of the normalized derivatives corresponds to the wavelength of the signal:

$$\sigma_{extremum} = \lambda \frac{\sqrt{\gamma m}}{2\pi} \tag{3.13}$$

where m is the derivative order and λ the wavelength. The corresponding expression for the second derivative centered at a one-dimensional Gaussian function is given in [69, 77]:

$$\sigma_{extremum} = \sigma_{Gaussian} \sqrt{\frac{\gamma}{3/2 - \gamma}} \tag{3.14}$$

Notice that if $\gamma = 3/4$, than $\sigma_{extremum} = \sigma_{Gaussian}$. For the Laplacian operator applied to the step-edge (cf. figure 3.8(f)) we obtain the following relation:

$$\sigma_{extremum} = x\sqrt{3 - 2\gamma} \tag{3.15}$$

Therefore, there is no extremum over scale for $\gamma \geq \frac{3}{2}$. The corresponding equation for the first derivative (cf. figure 3.8(c)) is:

$$\sigma_{extremum} = x\sqrt{\frac{1}{1 - \gamma}} \tag{3.16}$$

This shows that the first derivative can attain an extremum but the normalization factor must be $\gamma < 1$. The parameter γ can also take negative values but the lower the γ value the lower the magnitude of local extremum and it is therefore less distinctive. Note that these relations are valid for a perfect theoretical image structure. In the case of real images the texture, which is often present in the neighborhood of corners or edges, can change the scale trace. Nevertheless, the above relations show the ability of the differential expressions to select the scale of a local image structure and also show the influence of the normalization factor γ on the scale-space maxima. Experimental results verify these relations.

3.2.3 Differential expressions for scale selection

The derivatives computed in the Cartesian coordinates are generally not related to image structures, therefore useful structural operators are constructed from combinations of several Gaussian derivatives [38, 60, 120]. In this section we present the operators, which are often used in the context of scale selection of local features. A scale selection operator should be at least invariant to rotation to preserve a minimum of invariance. Illumination invariance is less critical because the features are localized at local extrema of the functions. However, one should be careful because the saturation can introduce an error. The localization of an extremum is independent of affine illumination changes, only the magnitude of the response changes. Scale selection using the gradient magnitude (cf. equation 3.17) has

also been used by Lindeberg in [69]. Chomat et al. [17] show that the gradient operator is appropriate for selecting the characteristic scale of local features and is robust to noise in the image.

$$\text{Squared gradient} \quad \sigma_D^2(L_x^2(\mathbf{x},\sigma_D) + L_y^2(\mathbf{x},\sigma_D)) \tag{3.17}$$

The magnitude of the gradient is naturally invariant to rotation and the phase can be used to determine the dominant orientation in the local feature (cf. section 5.2.2).

The Laplacian function is circularly symmetric and has been successfully used by Lindeberg [70] for blob detection and automatic scale selection.

$$\text{Laplacian} \quad \sigma_D^2|L_{xx}(\mathbf{x},\sigma_D) + L_{yy}(\mathbf{x},\sigma_D)| \tag{3.18}$$

The difference-of-Gaussian operator used by Lowe [73] is an approximation of the Laplacian-of-Gaussian and allows to accelerate the computation of a scale-space representation.

$$\text{Difference-of-Gaussian} \quad |I(\mathbf{x}) * g(\sigma_I) - I(\mathbf{x}) * g(k\sigma_I)| \tag{3.19}$$

A more sophisticated approach is to select the scale for which the trace and the determinant of the Hessian matrix assume a local extremum.

$$\max(|\text{trace}(\mathcal{H})|) \text{ and } \max(|\det(\mathcal{H})|) \tag{3.20}$$

Scale selection using the determinant of the Hessian matrix was used in [69]. The trace of the Hessian matrix (cf. equation 3.18) is equal to the Laplacian, but the simultaneous selection of the maxima of the determinant gives rise to points, for which the eigenvalues of the matrix have comparable and large values. These points are more robust against noise and illumination changes. Interest point detector, proposed by Moravec [90], improved by Harris [48] and later on by Schmid et al. [107], is based on the same idea, but it uses the components of the second moment matrix (cf. equation 3.2). A very similar detector was also developed by Förstner and Gülch [37].

$$\text{Harris function} \quad \det(\mu(\mathbf{x},\sigma_I,\sigma_D)) - \alpha\,\text{trace}^2(\mu(\mathbf{x},\sigma_I,\sigma_D)) \tag{3.21}$$

However, this operator was not adapted to scale changes. In order to deal with such transformation, Dufournaud et al. [30] parametrized the Harris operator by the scale. This enables interest points to be detected at different scales.

All the above operators have been used in the context of feature detection invariant to scale changes. However, no comparative evaluation has been presented in literature. In the next section we evaluate each of these functions.

3.2.4 Experimental evaluation

In the following we discuss the scale selection technique applied to real images and we present the results of an experimental evaluation of the operators presented in the previous section.

Characteristic scale. Given a point in an image and a scale selection function we compute the function responses for the set of scales σ_n (cf. figure 3.9). The characteristic

scale corresponds to the local extremum of the function. Note, that there might be several maxima or minima, therefore several characteristic scales. The characteristic scale is relatively independent of the image resolution. It is related to the structure and not to the resolution at which the structure is represented. The ratio of the scales, at which the extrema are found for corresponding points in two images is equal to the scale factor between the images. Instead of detecting extrema we can also look for another *easily recognizable* signal shape such as a zero-crossing of the second derivative. However, one should be careful with zero-crossings as the scale invariance of the response value is not necessary preserved for an arbitrary combination of derivatives. The maximum of the derivatives indicates the largest image content variation. The local shape of scale trace that we look for, depends on the order of the detection function and on the image structure we want to detect. For example, the zero crossing of the second derivative indicates the largest signal variation while the maximum indicates the distance to the signal changes and in consequence, the size of the structure. When the size of Laplacian kernel matches with the size of a blob-like structure the response attains an extremum. The Laplacian kernel can be also interpreted as a matching filter [29]. The Laplacian is well adapted to blob detection due to its circular symmetry, but other local structures such as corners, edges, ridges and multi-junctions are not rejected by this operator (cf. section 3.2.2). Many research results confirmed the usefulness of the Laplacian function for scale selection [17, 66, 71, 73].

Figure 3.9 presents an example of a real corner and edge. We can notice that the scale trace for the real corner and edge is similar to the scale trace of the model presented in figure 3.8. The method selecting scales corresponding to the extrema of both, the trace and the determinant of the Hessian matrix (cf. equation 3.20), may reject many feature if the scales are not the same. Notice that the scales corresponding to the maximum of the Hessian determinant and the Hessian trace are different for the real corner in figure 3.9. We can observe that the Harris measure (cf. equation 3.21), which is based on first derivatives, also attains a maximum over scale. The bottom rows of the frames present the scale trace for normalization factor $\gamma = 0.5$. The maxima over scale move to lower scales as explained in section 3.2.2.

In the case of a corner (cf. figure 3.8) the local features detected by gradient based methods [19, 48, 90] are not localized exactly at the junction of step edges. This is due to the evolution of interest points with respect to detection scale [3, 27]. The interest point is in fact localized inside the corner and its location changes in the gradient direction relatively to the detection scale. Moreover, the neighborhood of a corner is mostly textured in a real case. Thus, the second derivative (Laplacian) does not give zero response at these locations. As the experimental results show (cf. figure 3.11), the characteristic points are detected in the neighborhood of the signal changes and not exactly at the strongest signal variations. This happens for different detection functions, which are evaluated in the next paragraph. This permits the second derivative to be applied in order to measure the distance to the signal variations for features detected with first order derivative. The response attains a maximum when the scale of the second derivatives matches with the distance to the signal change. The invariant scaling property holds for any differential expression normalized with respect to scale. The local maxima over scales are preserved although the absolute value of the response may differ for transformed image patterns.

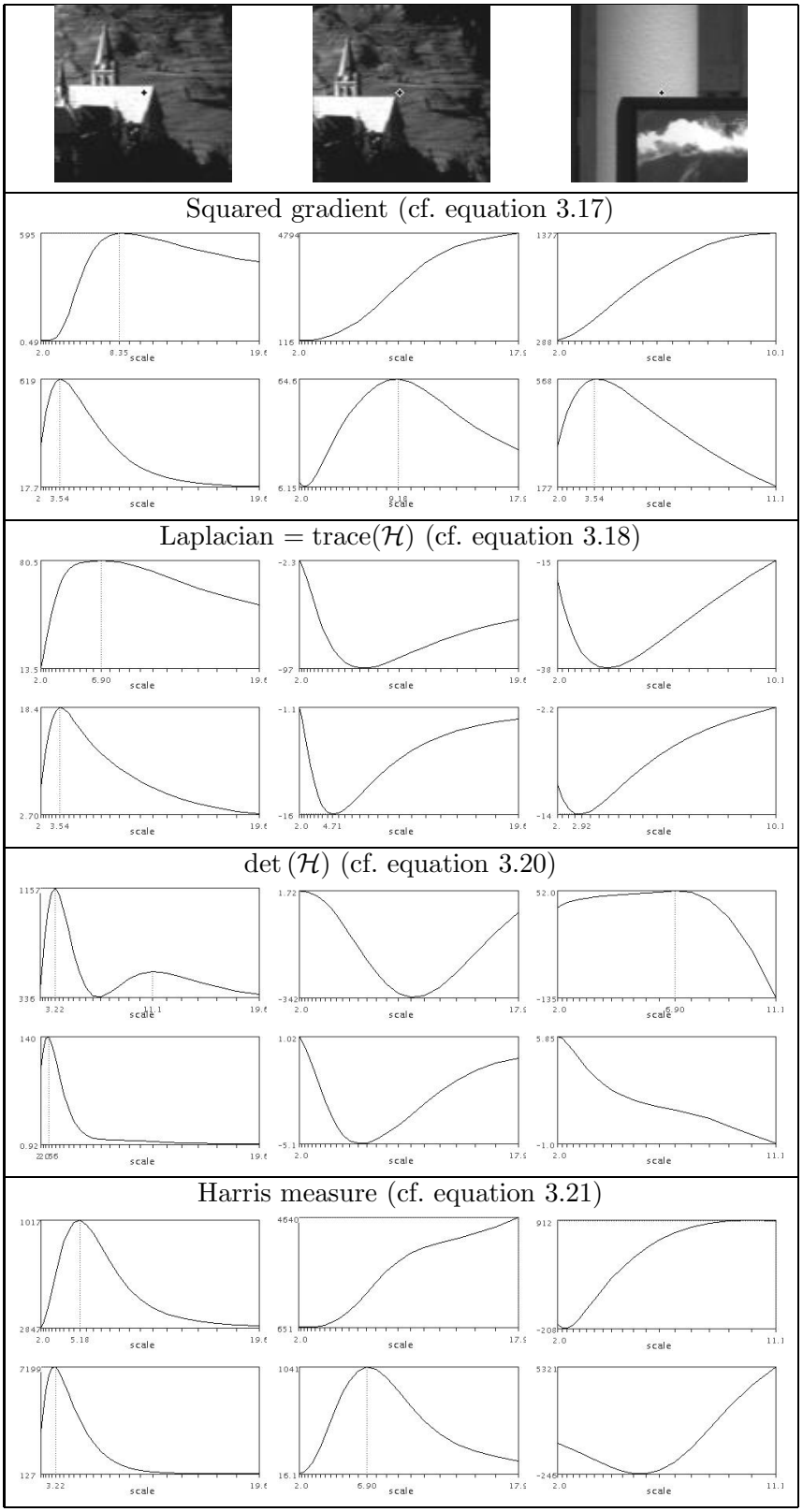


FIG. 3.9: *Scale trace of differential expressions applied to a real corner and edge. First column: point inside a corner. Second column: point outside a corner. Third column: point nearby an edge. In the frames: Top rows: Scale trace for* $\gamma = 1$. *Bottom rows: Scale trace for* $\gamma = 0.5$.

We have to consider the limited size of a neighborhood of interest points, while looking for a scale of local feature. It means that a high frequency content is more important in an observed region. On the other hand high frequencies are more sensitive to noise. The higher order derivatives enable more complex signal variations to be detected within a small region, but unfortunately, they can be easily influenced by a small localization error.

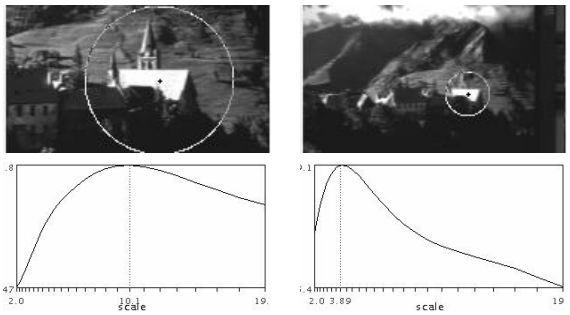


FIG. 3.10: *Example of characteristic scales. The top row shows two images taken with different focal lengths. The bottom row shows the response* $F_{norm}(\mathbf{x}, \sigma_n)$ *over scales where* F_{norm} *is the normalized Laplacian (cf. eq.3.18). The characteristic scales are at 10.1 and 3.89 for the left and right image, respectively. The ratio of scales corresponds to the scale factor (2.5) between the two images. The radius of displayed circles in the top row is equal* 3σ.

Evaluation results. The scale selection technique based on local maxima was evaluated for different differential operators given by equations (3.17–3.21). The evaluation was carried out on several sequences with scale changes up to a factor of 4.4. We computed 17 scale levels $\sigma = 1.2^n$, where $n = 0 \dots 16$ denotes the scale level. The scale selection technique was applied to every point in the image. Figure 3.11 illustrates the experiments for the Laplacian operator. It displays the image and the points for which the characteristic scale was selected (white and grey). Black points are the points for which the function (Laplacian) attains no maximum over scale. Note that these points lie in homogeneous regions and have no extremum in the range of applied scales. It might appear an extremum due to the noise but usually the amplitude of such extremum is very small and can be rejected by an arbitrary chosen threshold. Points with correctly selected scales are displayed in white. The selected scale is correct if the ratio between the characteristic scales in corresponding locations is equal to the scale factor between the images. The corresponding points are determined by projection with the estimated homography. In the case of multiple scale maxima, the point is considered correct, if one of the maxima corresponds to the correct ratio. If the ratio is different than the true scale change the selected scale is considered incorrect (in gray).

Our experiments show that the extrema are usually detected in the neighborhood of a significant signal change, where the gradient value is large. However, an extremum over scale is rarely attained exactly in the point where the gradient attains a local extremum in the image plane. Notice that the corner junctions and edges are displayed in black (cf. figure 3.11) that is the scale was not selected for these points. The range of scales that is

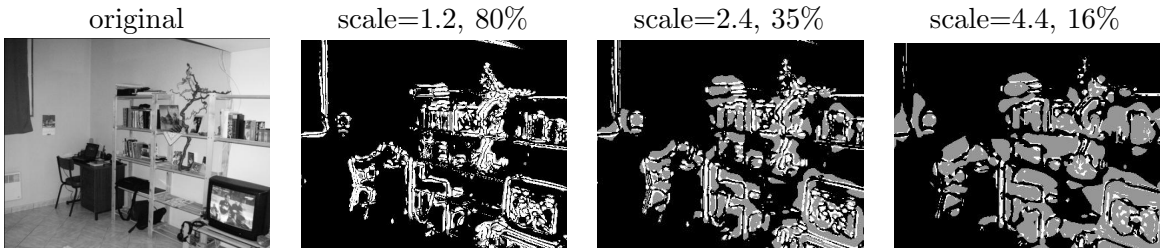


FIG. 3.11: *Characteristic scale of points selected by the LoG. Black – no characteristic scale is detected. Gray – a characteristic scale is detected. White – a characteristic scale is detected and is correct. The real scale change is given above the images and corresponds to* $scale = \frac{original}{scaled}$. *The percentage indicates the relative number of points for which the scale was correctly selected. The size of the images has been enlarged for the displaying purpose.*

explored in searching the extrema is limited and must be the same for all images, if we are given no prior knowledge about the scale factor between the images. Therefore, the larger the resolution change between images the fewer scale levels can be matched between the scale-space representations. The coarser scale limit is constrained by the finite size of the image. The finest scale is limited by the minimal size of the image structure that is considered to be a salient and detectable feature.

A problem appears if a characteristic scale is found near the outer scale for a point in the coarse resolution image. The characteristic scale of the corresponding point in the fine resolution image is then beyond the applied scale range, and vice-versa. Our experiments show that the performance of the method depends on the scale range, which determines the number of scale levels that match between two images. However, we cannot apply too wide a range of scales as we loose the local character of detected features, and the influence of image boundaries becomes too important. We can observe in figure 3.11 that there is only a small percentage of selected scales which are correct for large scale factors. There is therefore a need to select more prominent features, which exist at a wide range of scales.

F	det.%	corr. / det.%	cor. / tot.%	scales per point
Harris	16%	21%	3.4%	1.01
Lap.	46%	29%	13.3%	1.44
DoG	38%	28%	10.6%	1.38
grad.	30%	22%	6.6%	1.08
Hessian	17%	36%	6.1%	1.18

FIG. 3.12: *Scale selection results for points in images with scale change of factor 4.4. Column 1: Function applied for scale selection. Column 2: Percentage of points for which a characteristic scale is detected in the image. Column 3: Percentage of points for which a correct scale is detected with respect to detected scales. Column 4: Percentage of correct / total. Column 5: Average number of characteristic scales per point.*

In table 3.12 we have compared the results for different functions applied on images with a scale change of a factor 4.4. The results are averaged over several sequences. The first column indicates the functions applied for scale selection. The second one shows the percentage of points in the images for which a characteristic scale is detected. This parameter is less important but shows the selectivity of the detectors. We can observe that the most points are detected by the Laplacian. The percentage of correct points with respect to the number of detected points is given in column three. The best results have been obtained by the detector based on the eigenvalues of the Hessian matrix. As expected the LoG and the DoG has a similar and high score. A very important property of a feature detector is the overall number of reliable features detected in an image (correct/total). The fourth column shows the overall percentage of correct detection with respect to the number of points in the image. The number of useful features is an important measure of performance , as the image cannot be reliably represented by a few points. As we can see in table 3.12 this number is very different for the evaluated detectors. The largest number of correct points is detected by the Laplacian. The percentage is twice as high as for the gradient, and four times higher than for the Harris function. The LoG and the DoG results are again comparable. In the last column we show the average number of characteristic scales per point. The higher number of scales per point increases the probability that one of the scales will match with the scales selected for the corresponding point. On the other hand the probability of an accidental match and the complexity of the method is increased by a higher number of interest points. Once again the LoG and the DoG detector obtained the highest score.

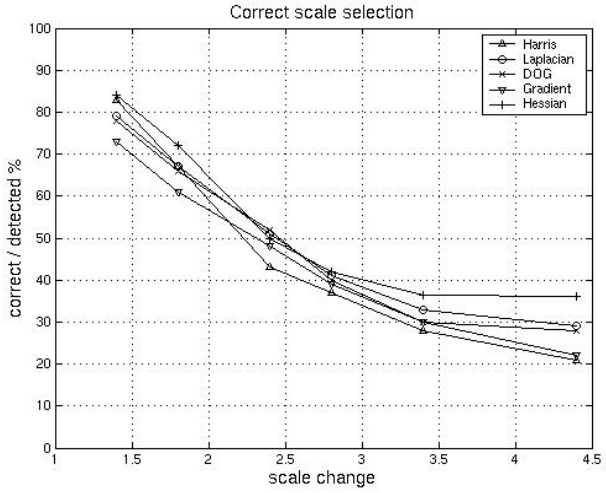


FIG. 3.13: *The percentage of correctly selected scales with respect to the detected characteristic scales.*

Figure 3.13 shows the percentage of scales correctly selected with respect to detected scales in image (correct/detected, column 3 in table 3.12). Note that the percentage does not differ very much for different detectors. This means that the detectors are equivalently affected by the limited range of scales and other negative effects, like noise, illumination

change etc.

3.3 Discussion

In this chapter we have presented the Gaussian scale-space representation and the mathematical background for invariant feature detection. Numerous research results have proven that the Gaussian function is the best tool to generate images at different resolutions. We base our approach on operators derived from this function.

We have introduced an isotropy measure based on the eigenvalues of the second moment matrix that can be used to estimate the affine deformation of an isotropic structure. The components of this matrix directly express the necessary affine transformation to obtain isotropic points and therefore the points, which covariantly change under arbitrary affine transformation. This property is explored to obtain an affine invariant detector, which is described in the next chapter of this manuscript.

We have presented a method for computing affine derivatives, which are useful for estimating the second moment matrix in the affine scale-space. We have shown that the normalized derivatives have comparable response for the same signal represented at different scales. Furthermore, we have shown that the maxima over scale of normalized derivatives correspond to characteristic scales of the local structure and are independent of image resolution.

It is important to select the best method for each part of the feature detection algorithm in order to obtain a high number of stable features. The scale selection technique based on the extrema of a scale-space representation is actually the most reliable method for determining the characteristic scale of a local structure. The experimental evaluation of automatic scale selection proves the usefulness of this technique and enables the Laplacian-of-Gaussians for our algorithm to be selected, as it gives the best results. Unlike the LoG operator, the Harris function does not give reliable scale selection results, which confirms the observation described in [7]. The theoretical analysis and the experimental evaluation provides a solid background for a new feature detection approach, which is detailed in the next chapter.

Chapter 4

Interest point detectors

INTEREST POINTS are characteristic points in an image, where the signal changes bi-dimensionally. They can be extracted reliably, are robust to partial visibility and the information content in these points is high. There are numerous approaches for interest point detection and the points extracted by these algorithms differ in localization, scale and structure (corners, blobs, multi-junctions). The average information content is different for points detected with different methods, which makes them more or less distinctive. In general, the objective is to develop a detector invariant to the most frequent geometric and photometric transformations introduced by different viewing conditions. A real scene usually contains locally smooth surfaces. A locally smooth surface can be approximated by piecewise planar surfaces. A planar surface undergoes perspective transformation if viewed from different viewpoints. The perspective deformation can be locally approximated by an affine transformation. As a conclusion, we assume that affine invariant features can reliably represent a locally smooth object.

In this chapter we propose a novel solution to detect scale and furthermore affine invariant features. The proposed detectors are the main contributions of the thesis. Our first approach is invariant to scale changes and is presented in section 4.1. This approach is extended in section 4.2 and is fully invariant to geometric and photometric affine transformations. It can be seen as a general solution to the affine problem of local feature detection. To evaluate and to compare the performance of detectors we propose the repeatability criterion that takes into account all the essential parameters defining local features. The comparative evaluation is presented in section 4.3.

4.1 Scale invariant detector

A scale invariant detector should be able to handle significant scale changes and can be used if weak affine transformations are expected. There are several methods proposed to

this problem in literature. These methods are reviewed in section 4.1.1. Our scale invariant detector is based on the Harris and the Laplacian function and is presented in section 4.1.2. The points extracted by this method are analyzed in section 4.1.3.

4.1.1 State of the art

In the following we present the existing methods, which deal with the problem of scale change. We first discuss the multi-scale approach and then the scale invariant detectors.

Multi-scale detector. Previous approaches to scale change problems suggest extracting points at several scales and using all these points to represent an image. In the approach proposed by Schmid and Mohr [107] the points are extracted with the Harris detector [48], which is invariant to image rotation. The Harris measure (cf. equation 3.21) enables the points to be selected, for which two eigenvalues of the autocorrelation matrix are large. This detector has shown to be the most robust against noise and illumination conditions, but fails in the presence of scale changes between images [108]. In order to deal with such a transformation Dufournaud et al. [30] proposed the scale adapted Harris operator. The points are detected at the local maxima of the Harris function applied at several scales. The normalized derivatives enable a comparable strength of cornerness to be obtained for points detected at different scales. Finally, a threshold enables the rejection of less significant corners. The scale adapted detector improves the repeatability (cf. section 4.3) of interest points. Given prior knowledge on the scale change between two images we can adapt the detector and extract the interest points present only at the selected scale. This provides us with points, for which the localization and scale perfectly reflect the real scale change between two images.

In general, the problem with a multi-scale approach is when a local image structure is present in a certain range of scales. The points are then detected at each scale within this range. As a consequence, there are many points, which represent the same structure, but the localization and the scale of the points is slightly different. The unnecessarily high number of points increases the probability of mismatches and the complexity of matching and recognition algorithm. Therefore, efficient methods for rejecting the false matches and for verifying the results are necessary at further steps of the algorithms.

Scale invariant detectors. There are very few approaches which are truly invariant to a considerable scale change. Notice that the affine invariance also includes the scale invariance, therefore the approaches related to scale invariance can also be found in section 4.2.1, which presents affine invariant detectors. However, the existing affine invariant detectors consider a very limited scale change. The results presented in literature concern the scale change, which does not exceed a factor of 2, while the standard Harris detector, not adapted to scale change, gives satisfying results approximately up to a scale factor of 1.4 [109]. We consider that a detector is scale invariant if it gives reliable results, at least, up to a scale factor of 4, which means that the size of an image structure changes with the same factor. The scale change is uniform, that is the same in every direction, although the detectors are robust to weak affine transformations.

Existing methods search for maxima in the 3D representation of an image (x, y and *scale*). This idea for detecting local features was introduced in early eighties by Crow-

ley [21, 22] and the pyramid representation was computed with difference-of-Gaussian filters. A feature point is detected if a local maximum is present in a surrounding 3D cube and if its absolute value is higher than a certain threshold. In figure 4.1 the point **m** is a feature point, if $F(\mathbf{m}, \sigma_n) > F(\bullet, \sigma_l)$ with $l \in \{n-1, n, n+1\}$ and $F(\mathbf{m}, \sigma_n) > threshold$. The existing approaches differ mainly in the differential expression used to build the scale-space representation.

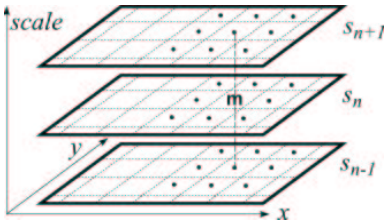


FIG. 4.1: *Searching for local maxima in scale-space representation.*

Lindeberg [70] searches for 3D maxima of the scale normalized Laplacian-of-Gaussian (LoG) function (cf. equation 3.18). The scale-space representation is built by successive smoothing of the high resolution image with Gaussian kernels of different size. This operator is circularly symmetric, therefore it is naturally invariant to rotation and well adapted for detecting blob-like structures. The experimental evaluation in section 3.2.4 shows the excellent ability of this function for the automatic scale selection. The scale invariance of interest point detectors with automatic scale selection has also been explored by Bretzner and Lindeberg [14].

Lowe [73] proposed an efficient algorithm for object recognition based on local 3D extrema in the scale-space pyramid built with difference-of-Gaussian (DoG) filters (cf. equation 3.19). The input image is smoothed with the Gaussian kernel of a fixed size. The smoothing is repeated a second time with the same filter. The first level of the difference-of-Gaussian representation is obtained by subtracting these two smoothed images. Next, the twice smoothed images are sampled with the scale factor corresponding to the scale of the kernel. The resulting sampled image is used to build the next DoG scale level. All the resolution levels are constructed by combined smoothing and sampling. The number of levels is limited by the image size. The local 3D extrema in the pyramid representation determine the localization and the scale of interest points. The DoG operator is a close approximation of the LoG function but this implementation permits a considerable acceleration of the computation process. A few images per second can be evaluated with this algorithm. Therefore it is an excellent tool for real time feature detection.

The common drawback of the DoG and the Laplacian representation is that the local maxima can also be detected in a neighborhood of contours or straight edges, where the signal change is only in one direction. These maxima are less stable because their localization is more sensitive to noise or small changes in neighboring texture. A more sophisticated approach, solving this problem, is to select the scale for which the trace and the determinant of the Hessian matrix simultaneously assume a local extremum (cf. equation 3.20). The trace of $\mathcal{H}$ matrix is equal to the Laplacian but detecting simultaneously the maxima of

the determinant gives rise to points, for which the second derivatives detect signal changes in two orthogonal directions. A similar idea is explored in the Harris detector, although it uses the first derivatives. The second derivative gives a small response exactly in the point where the signal change is most significant. Therefore the maxima are not localized exactly in the largest signal variation, but in the nearest neighborhood. The maxima nearby bi-dimensional change are more robust against noise and illumination conditions in comparison to the points detected by Laplacian or DoG interest points. Experimental evaluation (cf. section 3.2.4) showed that a local maximum of trace($\mathcal{H}$) at a point does not guarantee a maximum of det ($\mathcal{H}$) at the same point, and vice versa. Therefore fewer interest points are detected in an image comparing to Harris or DoG approach.

4.1.2 Harris-Laplace detector

In this section we propose a new interest point detector that combines the reliable Harris detector and the Laplacian based scale selection. The evaluation of interest point detectors presented in [109] showed the superiority of the Harris detector compared to other existing approaches [19, 36, 49, 50]. In our experiments (cf. section 3.2.4) we noticed that the scale adapted Harris function rarely attains maxima over scales in a scale-space representation. If too few interest points are detected, the image is not reliably represented. Therefore, we abandoned the idea of searching 3D maxima of the Harris function. Furthermore, the experiments showed that LoG function enables the highest percentage of correct characteristic scales to be found. Therefore, we propose to use the Laplacian to select the scales for points extracted with the Harris detector. Harris-Laplace detector uses the Harris function (cf. equation 4.1) to localize points in each level of the scale-space representation. Next, it selects the points, for which the Laplacian-of-Gaussian (cf. equation 4.2) attains a maximum over scale. In this way we combine these two methods to obtain a reliable interest point detector invariant to significant scale changes.

In the following we explain in detail the detection algorithm. We propose two implementations of the general idea presented above. The first one is a fast algorithm for detecting the location of interest points and the scale of associated characteristic regions. The second one renders an estimation of the exact location and scale of each interest point possible.

Harris-Laplace. Our detection algorithm works as follows. We first build a scale-space representation with the Harris function for arbitrary selected scales $\sigma_n = s^n\sigma_0$, where s is a scale factor between successive levels. At each level of the representation we extract the interest points by detecting the local maxima in the 8-neighborhood of a point $\mathbf{x}$. A threshold is used to reject the maxima of small cornerness, as they are less stable under arbitrary viewing conditions:

$$\det(\mu(\mathbf{x}, \sigma_\mathbf{n})) - \alpha \operatorname{trace}^2(\mu(\mathbf{x}, \sigma_\mathbf{n})) > threshold_H \tag{4.1}$$

The matrix $\mu(\mathbf{x}, \sigma_n)$ is in fact computed with the integration scale $\sigma_I = \sigma_n$ and the local scale $\sigma_D = k\sigma_n$, where k is a constant factor. In order to obtain a compact and representative set of points, we verify for each of the candidate points found on different levels whether it forms a maximum in the scale dimension $F(\mathbf{x}, \sigma_n) > F(\mathbf{x}, \sigma_l)$ with $l \in$

$\{n-1, n+1\}$ and $F(\mathbf{x}, \sigma_n) > threshold$. The Laplacian-of-Gaussian is used for finding the maxima over scale (cf. equation 4.2). We reject the points for which the Laplacian attains no extremum or its response is below a threshold.

$$\sigma_n^2 |L_{xx}(\mathbf{x}, \sigma_n) + L_{yy}(\mathbf{x}, \sigma_n)| > threshold_L \tag{4.2}$$

Extended Harris-Laplace. For some points the scale trace maximum does not correspond to the arbitrary set detection scales. These points are either rejected, due to the lack of the maximum, or the location and the scale are not accurate. The Harris-Laplace algorithm can be extended to find the location $\mathbf{x}$ and the scale σ_I of an interest point with a high accuracy. The detector can be initialized with multi-scale Harris point. Next, for each point we can apply an iterative algorithm that simultaneously detects the location and the scale of the points. A straightforward iterative method for feature detection can be expressed as follows. For a given initial point $\mathbf{x}$ with scale σ_I:

1. find the local extremum over scale for the point $\mathbf{x}^{(k)}$, otherwise reject the point. The range of scales can be limited by $\sigma_I^{(k+1)} = s\sigma_I^{(k)}$ with $s \in [0.7, \ldots, 1.4]$,
2. detect spatial location $\mathbf{x}^{(k+1)}$ of a maximum of the Harris measure (cf. equation 4.1) nearest to $\mathbf{x}^{(k)}$ for selected σ_I^{k+1},
3. go to step 1 if $\sigma_I^{(k+1)} \neq \sigma_I^{(k)}$ or $\mathbf{x}^{(k+1)} \neq \mathbf{x}^{(k)}$.

The initial points can be detected with larger scale change between two successive representation levels, i.e. $s = 1.4$. A smaller scale interval i.e. $s = 1.12$, in the iterative algorithm gives better approximation of the location and scale. As we can imagine the initial points detected on the same local structure but at different representation levels should converge to the same location and the same scale. It is straightforward to find the similar points using point coordinates and scales. To represent the structure we can keep only one of them. This approach provides us with points, for which the location and scale are estimated with a high accuracy. It also finds correct parameters for points, which are rejected by the Laplacian measure in the Harris-Laplace approach. However, the iterative algorithm applied for each initial point is more time consuming in comparison to the Harris-Laplace approach.

4.1.3 Scale covariant points

In figure 4.2 we present several examples of points detected with the Harris-Laplace method. The top row shows points detected with the multi-scale Harris detector. The detection scale is represented by a circle around a point with radius $3\sigma_I$. Note, how the interest point, which is detected for the same image structure, changes its location in the gradient direction relative to the detection scale. One could determine the chain of points and select only one of them to represent the local structure [3] (cf. figure 4.4). The similar points are located in a small neighborhood and can be determined by comparing their descriptors. However, for the local structures existing in a wide range of scales the information content can change. In our approach the LoG measure is used to select

the representative points for such structures. Moreover, the Laplacian enables the corresponding characteristic points to be selected (bottom row) even if the transformation between images is significant. Sometimes, two or more points are selected, but given no prior knowledge about the scale change between images we have to keep all the selected points. The location and the scale of points is correct with respect to the transformation between images.

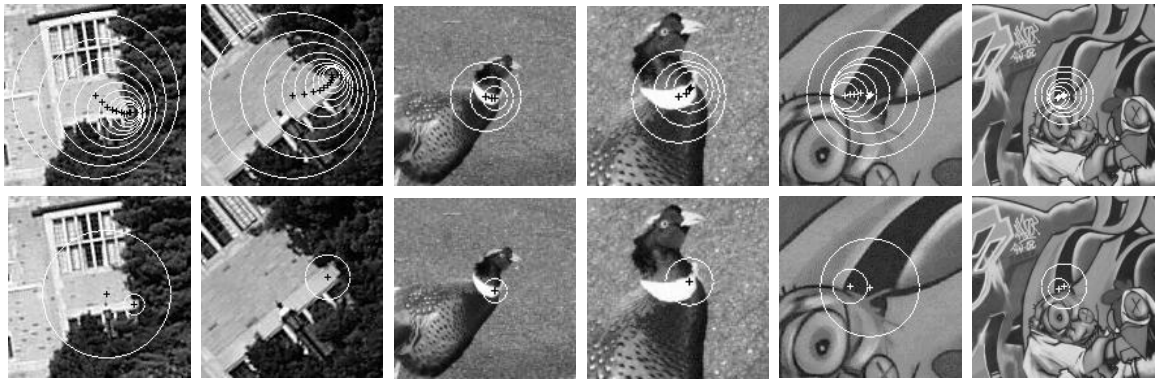


FIG. 4.2: *Scale invariant interest point detection: (top) Initial multi-scale Harris points (bottom) Selected points with the Laplacian measure.*

Figure 4.3 shows the scale-space representation for two images with points detected by the Harris-Laplace method. For each scale level of the object we present the selected points. There are many point-to-point correspondences between the levels for which the scale ratio corresponds to the real scale change between the images (indicated by pointers). Additionally, very few points are detected in the same location but on different levels. Our points are therefore characteristic in the image plane and in the scale dimension.

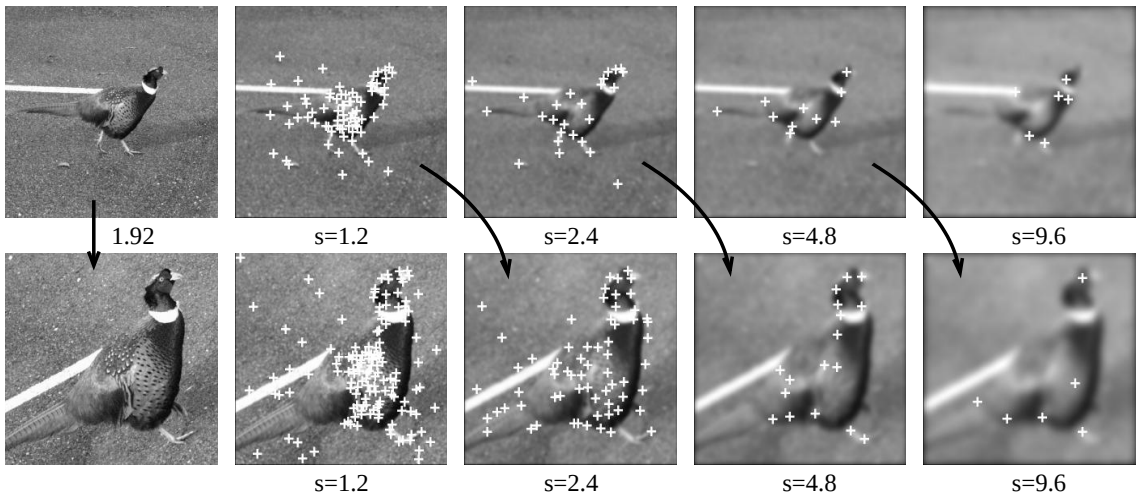


FIG. 4.3: *Points detected on different resolution levels with the Harris-Laplace method. The detection scale is given below the images*

4.2 Affine invariant detector

An affine invariant detector can be seen as a generalization of the scale invariant detector. In the case of affine transformation the scaling can be non-uniform, that is different in each direction. The non-uniform scaling has an influence on the localization, the scale and the shape of characteristic local structures. Therefore, the scale invariant detectors fail in the case of significant affine transformations. In section 4.2.1, we briefly review the existing approaches related to the affine invariant detection. Next, in section 4.2.2, we explain in detail the consecutive steps of the extraction method. The detection algorithm is followed by the analysis of the affine covariant points (cf. section 4.2.3).

4.2.1 State of the art

An affine invariant algorithm for corner detection was proposed by Alvarez and Morales [3]. They apply affine morphological multi-scale analysis to extract corners. The corner is represented by a local extremum of the response of the differential operator:

$$L_y^2 L_{xx} - 2L_x L_y L_{xy} + L_x^2 L_{yy} > threshold$$

which corresponds to the second derivative in the direction orthogonal to the gradient. By an iterative procedure, for each extracted point (x_0, y_0, σ_0) they build a chain of points (x_n, y_n, σ_n) associated with the same local image structure. They assume that the evolution of a corner is given by a straight line formed by points (d_n, σ_n) where d_n is a distance between the point (x_n, y_n, σ_n) and the point (x_0, y_0, σ_0) (cf. figure 4.4). The slope of this line identifies the angle of the corner. The point moves along the bisector line of the corner. This assumption together with the recovered angle enables the initial location and orientation of the corner in the image to be computed. Similar idea was previously explored by Deriche and Giraudon [27]. The main drawback of this approach is that an interest point in images of natural scenes cannot be approximated by a model of a perfect corner, as it can take any form of a bi-directional signal change. The real points detected at different scales do not move along a straight bisector line as the texture around the points significantly influences their location. This approach cannot be a general solution to the problem of affine invariance but can give good results for synthetic images where the corners and multi-junctions are formed by straight or nearly straight step-edges.

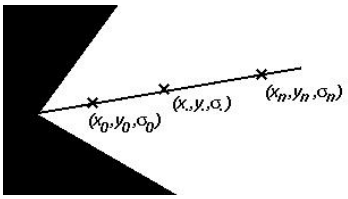
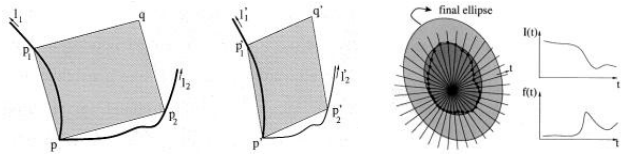


FIG. 4.4: *Chain of interest points* (x_n, y_n, σ_n) *associated with the same corner.*

Recently, Tuytelaars and Van Gool [127, 128] proposed two approaches for detecting image features in an affine invariant way. The first one starts from corners and uses the

nearby edges (cf. figure 4.5(a)). The first step is the extraction of Harris points, which limits the search regions and reduces the complexity of the method. Two nearby edges, which are required for each point, additionally limit the number of potential features in an image. One point moving along each of the two edges together with the Harris point determine a parallelogram. The points stop at positions where some photometric quantities of the texture covered by the parallelogram go through an extremum. Several intensity based functions are used for parallelogram evaluation. In this approach a reliable algorithm for extracting the edges is necessary. This method looks for a specific structure in images therefore we can categorize it as a model based approach. The second method is purely intensity-based and starts with extraction of local intensity extrema. Next, they investigate the intensity profiles along rays going out of the local extremum. A marker is placed on each ray in the place, where the intensity profile significantly changes. Finally, an ellipse is fitted to the region determined by the markers (cf. figure 4.5(b)).



(a) (b)

FIG. 4.5: *a) Regions determined by Harris points and edges. b) Intensity profiles and an ellipse fitted to the profiles.*

Lindeberg and Garding [71] developed a method to find blob-like affine features using an iterative scheme, in the context of shape from texture. The affine invariance of shape adapted fixed points was also used for estimating surface orientation from binocular data (shape from disparity gradients). This work provided a theoretical background for our affine invariant detector described in the next section. The algorithm explores the properties of the second moment matrix described in section 3.1.3 and iteratively estimates the affine transformation of local patterns. The method was used to recover the surface orientation. The authors propose to extract the points using the maxima of a uniform scale-space representation and to iteratively modify the scale and the shape of points. However, the location of points is detected only at the initial step of the algorithm, by the circularly symmetric, not affine invariant Laplacian measure. Therefore, the spatial location of the maximum can be slightly different if the pattern undergoes an affine deformation. The method was applied to detect elliptical blobs in the context of hand tracking [63]. This approach was implemented in the domain of matching and recognition by Baumberg [7]. He extracts interest points at several scales using the Harris detector and then adapts the shape of the regions to the local image structure using the iterative procedure proposed by Lindeberg. This enables affine invariant descriptors to be obtained for a given fixed scale and fixed location, that is the scale and the location of the points are not extracted in an affine invariant way. The points as well as the associated regions are therefore not invariant in the case of significant affine transformations (see section 4.3 for a quantitative comparison). Furthermore, the multi-scale Harris detector extracts many points which are

repeated at the neighboring scale levels (cf. figure 4.2). As a result there are many points representing the same corners. This increases the probability of false matches and in the case of indexing the complexity is increased.

Recently, Schaffalitzky and Zisserman [104] extended the Harris-Laplace detector by affine normalization with the algorithm proposed by Baumberg. This detector suffers from the same drawbacks as the location and scale of points are not extracted in an affine invariant way, although the uniform scale changes between the views are handled by the scale invariant Harris-Laplace detector. The affine normalization technique was also used by Schaffalitzky and Zisserman [103] for affine rectification of textured regions. The affine invariant descriptor enabled the corresponding textured regions to be determined. Next, the interest points were extracted by the Harris detector within the affine normalized regions and used to verify the matching correctness. This technique requires an initial segmentation to extract the textured regions.

4.2.2 Harris-Affine detector

In the case of affine transformations the scale change can be different in each direction. The presented Harris-Laplace detector will fail in the case of important affine transformation because it assumes a uniform scale change. Figure 4.8 presents two pairs of points detected in images with a significant affine deformation. The top row shows points detected with the multi-scale Harris detector. The scale, selected with the Laplacian, is displayed in black. If we project the circular neighborhood of the corresponding point using the affine transformation, we obtain an elliptical region which does not cover the same part of the image. We can see the projected points displayed in the bottom row (in white) superposed on the corresponding Harris-Laplace point (in black).

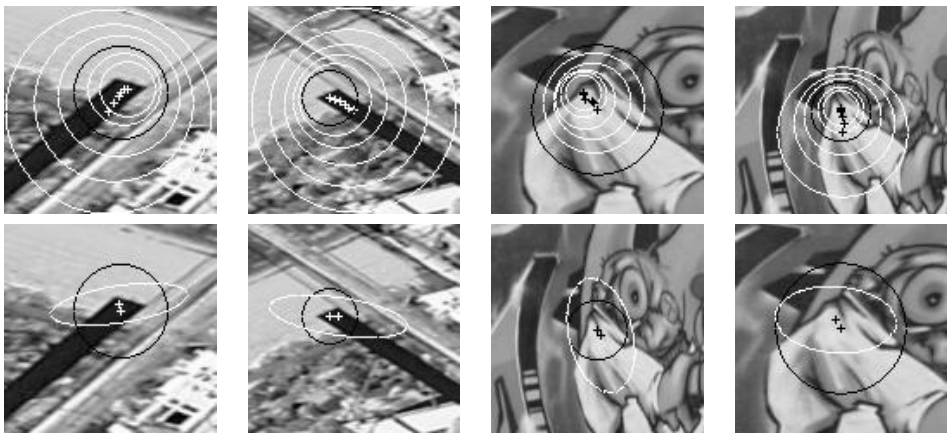


FIG. 4.6: *Non adapted interest point detection in affine transformed images: (Top) Initial interest points detected with the multi-scale Harris detector and their characteristic scales selected by Laplacian scale peak (in black – Harris-Laplace). (Bottom) Characteristic point detected with Harris-Laplace (in black) and corresponding point from the other image projected with affine transformation (in white).*

In the case of affine transformation, when the scale change is not necessarily the same in every direction, automatically selected scales do not reflect the real transformation of a point. It is well known that the local Harris maxima change the spatial location with respect to the detection scale (cf. figures 4.2 and 4.8). Thus, an additional error is introduced to the location of points if the detection scales do not correspond to the scale factors between corresponding image patterns. The detection scales in orthogonal directions have to vary independently, in order to deal with possible affine scaling. Suppose both scales can be adapted to a local image structure. Hence, we face the problem of computing second moment matrix in affine Gaussian scale-space, where a circular point neighborhood is replaced by an ellipse.

Numerous successful results in estimating affine deformation with the second moment matrix [2, 7, 71, 103, 104] proved the usefulness of this matrix. We explore its properties in order to select the detection scales. The adequate properties are described in section 3.1.3. For a given point $\mathbf{x}$ the second moment matrix μ in non-uniform scale-space is defined by:

$$\mu(\mathbf{x}, \Sigma_I, \Sigma_D) = g(\Sigma_I) * ((\nabla L)(\mathbf{x}, \Sigma_D)(\nabla L)(\mathbf{x}, \Sigma_D)^T)$$

where Σ_I and Σ_D are the covariance matrices which determine the integration and the differentiation Gaussian kernel. To reduce the search space we impose the condition $\Sigma_I = s\Sigma_D$, where s is a scalar. Furthermore, to limit the search space we initialize the affine detector with interest points extracted by multi-scale Harris detector. Any detector can be used to determine the *spatial localization* of initial points but the Harris detector is also based on the second moment matrix, thus it naturally fits in this framework. To obtain the *shape matrix* for each interest point we compute the second moment descriptor with automatically selected *integration* and *differentiation* scales. The outline of our detection method is presented in the following:

- the *spatial localization* of an interest point at a given scale and shape is determined by the local maximum of the Harris function,
- the *integration scale* is selected at the extrema over scale of normalized derivatives,
- the *differentiation scale* is selected at the maximum of normalized isotropy,
- the *shape adaptation matrix* normalizes the point neighborhood.

In the following we discuss in detail each step of the algorithm.

Shape adaptation matrix. Our iterative shape adaptation method works in the transformed image domain. As presented in section 3.1, instead of applying the affine Gaussian kernel we transform the image and apply a uniform kernel. That enables the use of the recursive implementation of uniform Gaussian filters for computing L_x and L_y. The second moment matrix is computed according to equation 3.2. A local window W is centered at interest point $\mathbf{x}$ and transformed by matrix:

$$U^{(k-1)} = (\mu^{-\frac{1}{2}})^{(k-1)} \cdots (\mu^{-\frac{1}{2}})^{(1)} \cdot U^{(0)} \tag{4.3}$$

in step (k) of the iterative algorithm. In the following we refer to this operation as U-transformation. Note, that a new μ matrix is computed at each iteration and the U matrix is the concatenation of square roots of the second moment matrices. We ensure that the original image is correctly sampled by setting the larger eigenvalue $\lambda_{max}(U) = 1$. It means that the image patch is enlarged in the direction of $\lambda_{min}(U)$. For a given point the integration and the local scales determine the second moment matrix μ. These scale parameters are automatically detected in each iteration step. Thus, the resulting μ matrix is independent of the initial scale and the resolution of the image.

Integration scale. For a given spatial point we automatically select its characteristic scale. In order to preserve the invariance to size changes we select the integration scale σ_I for which the normalized Laplacian (cf. equation 4.2) attains a local maximum over scale. In the case of weak scale changes it is sufficient to keep σ_I constant during the iterations. In the presence of important affine deformations the scale change is very different in each direction. Thus, the characteristic scale detected in the original image and in its U-transformed version can be significantly different. Therefore, it is essential to select the integration scale after applying the U transformation. We use a procedure similar to the one described for extended version of Harris-Laplace detector, in section 4.1.2. This allows the initial points to converge toward a point where the scale and the second moment matrix do not change any more. Note, that the extremum over scale has to be consequently of the same type during the iterations. Otherwise, the method can switch between a maximum and a minimum if there are both types of extrema in the scanned range of scales.

Differentiation scale. The local differentiation scale is less critical and can be set proportional to the integration scale $\sigma_D = s\sigma_I$, where s is a constant factor. However, we propose to base the derivative scale on the isotropy measure introduced in section 3.1.3. Factor s is commonly chosen from the range $[0.5, \ldots, 0.75]$. Our solution is to select the differentiation scale for which the local isotropy assumes a maximum over this range of scales. Given the integration scale σ_I we select $s \in [0.5, \ldots, 0.75]$ for which the $\mathcal{Q}$ measure (cf. equation 3.8) assumes a maximum. This solution is motivated by the fact that the local scale has an important influence on the convergence of the second moment matrix. The iterative procedure converges toward a matrix with equal eigenvalues. The smaller the difference between the eigenvalues $(\lambda_{max}(\mu), \lambda_{min}(\mu))$ of initial matrix, the closer the final solution is and the procedure converges faster. Note that the Harris measure (cf. equation 4.1) already selects the points with two large eigenvalues. A large difference between the eigenvalues leads to a large scaling in one direction by the U-transformation. The point does not converge to a stable solution due to noise. The selection of the local scale enables a reasonable eigenvalue ratio to be obtained and the points to converge, which would not converge if the ratio is too large.

Spatial localization. We have already shown how the local maxima of the Harris measure (cf. equation 4.1) change their location if the detection scale changes (cf. figure 4.8). We can also observe this effect, when the scale change is different in each direction. The detection with different scales in x and y directions is replaced by adapting the image and then applying the same scale in both directions. The affine normalization of a point neighborhood slightly displaces spatial maxima of the Harris function. Consequently, we re-detect the maximum in the affine normalized window W. Thus, we obtain a vector of

displacement to the nearest maximum in the U-normalized image domain. The location of the initial point is corrected with the displacement vector back-transformed to the original image domain:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + U^{(k-1)} \cdot (\mathbf{x}_w^{(k)} - \mathbf{x}_w^{(k-1)})$$

where $\mathbf{x}_w$ is the point in the coordinates of the U-transformed image.

Convergence criterion. The important part of the iteration procedure is the stopping criterion. The convergence measure can be based on either the U or the μ matrix. If the criterion is based on μ computed in each iteration step, we require that this matrix be sufficiently close to a pure rotation. This implies that $\lambda_{max}(\mu)$ and $\lambda_{min}(\mu)$ are equal. In practice we allow for a small error $\epsilon_C = 0.05$.

$$\frac{\lambda_{max}(\mu) - \lambda_{min}(\mu)}{\lambda_{max}(\mu)} < \epsilon_C \tag{4.4}$$

Another possibility is to decompose the matrix $U = R^T \cdot D \cdot R$ into rotation R and scaling D and compare the consecutive transformations. We allow the point if the consecutive R and D transformations are sufficiently similar. Both termination criteria give the same final results. Another important point is to stop the iteration in the case of divergence. In theory there is a singular case when the eigenvalue ratio tends to infinity. Therefore, the point should be rejected if the ratio is too large (i.e. $\epsilon_l = 6$), otherwise it leads to unstable elongated structures.

$$\frac{\lambda_{max}(D)}{\lambda_{min}(D)} > \epsilon_l \tag{4.5}$$

The convergence properties of the shape adaptation algorithm are extensively studied in [71]. It is shown that besides the singular case, the point of convergence is always unique. In general the procedure converges provided that the initial estimation of affine deformation is sufficiently close to the true deformation and the integration scale is correctly selected with respect to the size of the local image structure.

Detection algorithm. We propose an iterative procedure that allows the initial points to converge to affine covariant points, that is the points, which covariantly change with viewpoint. To initialize our algorithm we use points extracted by the multi-scale Harris detector. These points are not detected in an affine invariant way due to a non adapted Gaussian kernel, but provide an approximate localization and scale for further search for affine covariant interest points. For a given initial interest point $\mathbf{x}^{(0)}$ we apply the following procedure:

1. initialize $U^{(0)}$ to the identity matrix
2. normalize window $W(\mathbf{x}_w) = I(\mathbf{x})$ centered in $U^{(k-1)}\mathbf{x}_w^{(k-1)} = \mathbf{x}^{(k-1)}$
3. select *integration scale* σ_I in $\mathbf{x}_w^{(k-1)}$
4. select *differentiation scale* $\sigma_D = s\sigma_I$, which maximizes $\frac{\lambda_{min}(\mu)}{\lambda_{max}(\mu)}$, with $s \in [0.5, \ldots, 0.75]$ and $\mu = \mu(\mathbf{x}_w{}^{(k-1)}, \sigma_D, \sigma_I)$

5. detect *spatial localization* $\mathbf{x}_w^{(k)}$ of a maximum of the Harris measure (cf. equation 4.1) nearest to $\mathbf{x}_w^{(k-1)}$ and compute the location of interest point $\mathbf{x}^{(k)}$
6. compute $\mu_i^{(k)} = \mu^{-\frac{1}{2}}(\mathbf{x}_w^{(k)}, \sigma_D, \sigma_I)$
7. concatenate transformation $U^{(k)} = \mu_i^{(k)} \cdot U^{(k-1)}$ and normalize $U^{(k)}$ to $\lambda_{max}(U^{(k)}) = 1$
8. go to step 2 if $(1 - \lambda_{min}(\mu_i^{(k)})/\lambda_{max}(\mu_i^{(k)})) > \epsilon_C$

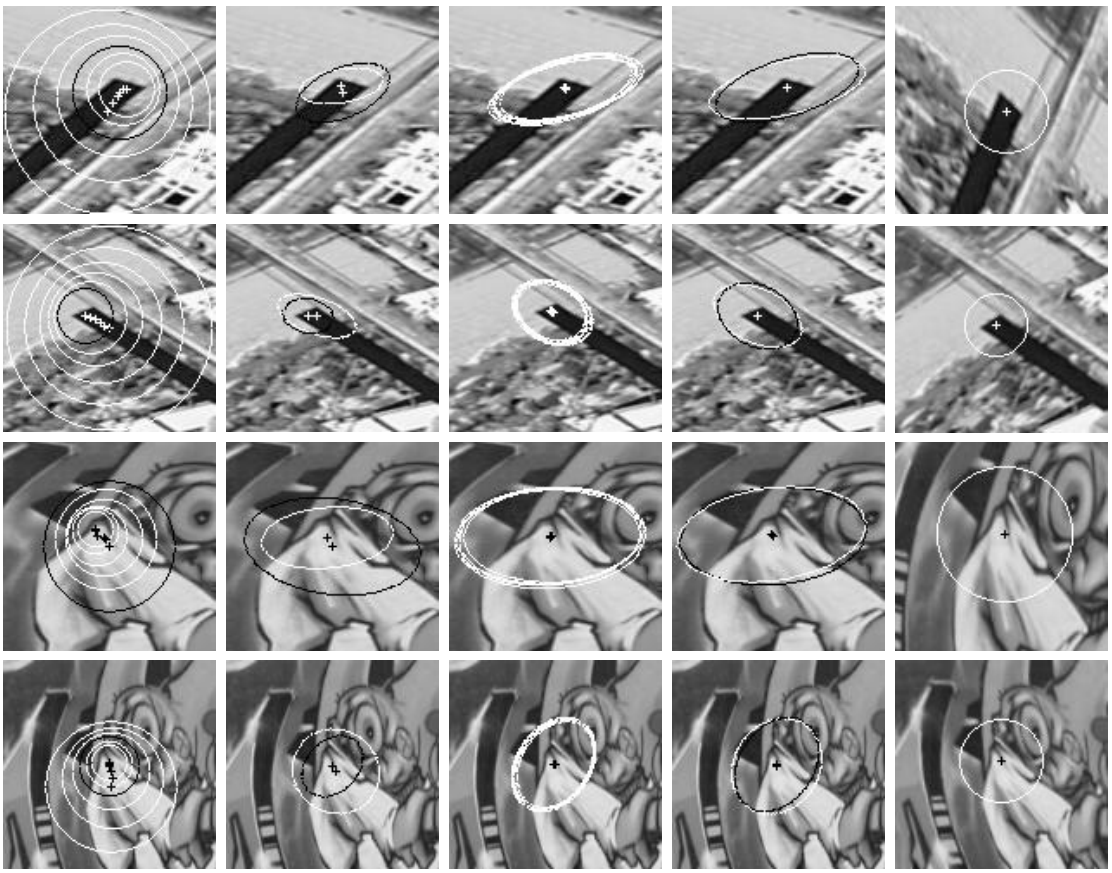
Although the computation may seem to be very time consuming, note that most time is spent on computing L_x and L_y, which is done only once in each step if the relation between the integration and local scales is constant. The iteration loop begins with selecting the integration scale because we have noticed that this part of the algorithm is most robust to small localization errors of an interest point. However, scale σ_I changes if the shape of the patch is transformed. Given an initial approximate solution, the presented algorithm enables one to iteratively modify the shape, the scale and the spatial location of a point and converges to a local structure, which is determined despite arbitrary affine transformations. Figure 4.7 shows affine points detected in consecutive steps of the iterative procedure. After the fourth iteration the location, scale and shape of the point do not change any more. We can notice that the elipses cover the same image region despite strong affine deformation.

FIG. 4.7: *Iterative evolution of affine interest point. In columns: initial points and successive approximations of localization, scale and affine deformation.*

Selection of similar affine points. Provided that the normalized region is isotropic, there is one spatial maximum of the Harris measure and one characteristic scale for the considered local structure. Therefore, several initial points corresponding to the same feature but detected at different scale levels can converge toward one point location and scale. It is straightforward to identify these points by comparing their location (x, y), scale σ_I, stretch $\lambda_{min}(U)$ and skew. The skew is recovered from the rotation matrix R, where $U = R^T \cdot D \cdot R$. We define a point similar if each of these parameters is reasonably close

to the parameters of the reference point. Finally, we compute the average parameters and select the most similar point from the identified set of points. As a result, for a given image we obtain a set of points, where each one represents a different image location and structure.

4.2.3 Affine covariant points



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

FIG. 4.8: *Affine invariant interest point detection : (a) Initial interest points detected with the multi-scale Harris detector and their characteristic scale selected by Laplacian scale peak (in black – Harris-Laplace). (b) Affine region detected for Harris-Laplace point (in black) and its corresponding point projected from the other image (in white). (c) Points and corresponding affine regions obtained after applying our iterative algorithm to initial multi-scale Harris points. (d) Selected* average *affine point (in black) and its corresponding projected point (in white). (e) Point neighborhoods normalized with the estimated matrices to remove stretch and skew.*

Figure 4.8 presents two examples of characteristic local structures. Column (a) displays the points used for initialization, which are detected by the multi-scale Harris detector.

The circle around the point shows the scale of detection, where the radius of the circle is $3\sigma_I$. The circles in black show the points selected by Harris-Laplace detector. Note that there is an important displacement between points detected at different scales and the circles in corresponding images (top and bottom row) do not cover the same part of the image. In column (b) we show the points (in black) detected by applying the iterative procedure to the Harris-Laplace points [7, 104]. The scale and the location of point is constant during iterations. The projected corresponding regions are displayed in white and clearly show the difference in localization and region shape. The initial scale is not correctly detected due to not adapted uniform Laplacian operator. Similarly, the location of points differs in 3-4 pixels. In our approach the points, which correspond to the same physical structure, but are detected at different locations due to scale, converge to the same point location. The effective number of interest points is therefore reduced. The affine covariant points, to which the initial points converge are presented in column (c). These points have been obtained by applying the algorithm described in the previous section. We can see that the method converges correctly even if the location and the scale of initial point is relatively far from the point of convergence. Convergence is in general obtained in less than 10 iterations. Further information on the statistics of convergence can be found in section 6.1.2

The minor differences between the regions in column (d) are caused by the imprecision of the scale estimation and the error ϵ_C. Column (e) shows the "average" points normalized with estimated matrices to remove the stretch and the skew. We can see clearly that the regions correspond between the two images (top and bottom row).

4.3 Comparative evaluation of detectors

In the previous sections we have presented several existing approaches and we have proposed two new solutions for scale and for affine invariant detection of interest points. In this section we present a qualitative and quantitative evaluation of these detectors. The stability of detectors is evaluated using the repeatability criteria introduced in [108]. The repeatability score for a given pair of corresponding images is computed as a ratio between the number of point-to-point correspondences that can be established and the number of detected points in the coarse scale image. In general, there are less points detected in the coarse scale image so the probability to find a match in the other image is higher. We consider two points $\mathbf{x}_a$ and $\mathbf{x}_b$ corresponding if:

1. the error in relative location of points $\|\mathbf{x}_a - H \cdot \mathbf{x}_b\| < 1.5$ pixel, where H is the homography between images,

2. the error in image surface covered by point neighborhoods is less than 20% $\epsilon_S < 0.2$ (cf. equations 4.6 and 4.7).

The location error of 1.5 pixel can be neglected because it is relatively small compared to the error introduced by the imprecision of the scale estimation. This criterion is rather restrictive as large scale features will have more pixels of imprecision. We take into account only the points located in the part of the scene present in both images.

4.3.1 Scale invariant detectors

In the following we compute the repeatability score for the scale invariant detectors. We compare the detection methods proposed by Lindeberg [71] (Laplacian, Hessian and gradient), Lowe [73] (DoG) as well as our Harris-Laplace and Harris-Affine detector. To show the gain compared to the non-scale invariant method, we also present the results for the standard Harris detector. Figure 4.9 shows the repeatability score for the compared methods. 20% surface error corresponds to 10% difference between the real scale change and the ratio of scales selected for corresponding points:

$$\epsilon_S = |1 - s^2 \frac{\sigma_a^2}{\sigma_b^2}| < 0.2 \tag{4.6}$$

where s is the real scale factor, which is recovered from the homography between the images. The experiments have been done on 10 sequences of real images (cf. annex A.3). We used planar scenes in order to apply the homography for verification. For 3D scenes the camera position was fixed and only the focal settings changed. Each sequence consists of scaled and rotated images, for which the scale factor varies from 1.4 up to 4.5.

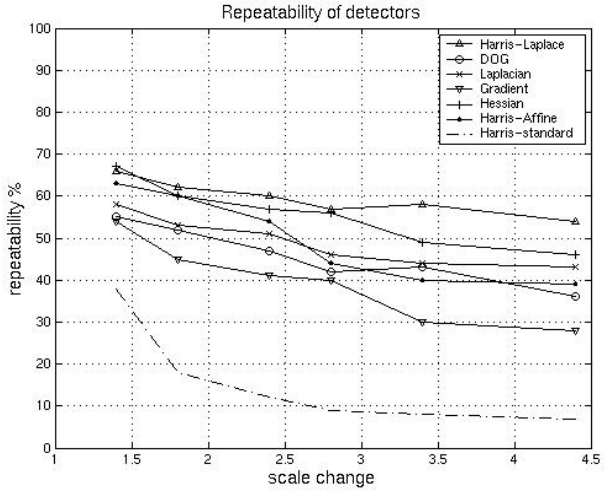


FIG. 4.9: *Repeatability of interest point detectors with respect to scale changes.*

Best results are obtained for the Harris-Laplace method. The best repeatability score is 68% for scale factor of 1.4. The repeatability is not 100% because some points cannot be matched. This is due to the range of applied scales, which is the same for both images (cf. section 4.1.2). The unmatched points are extracted from finer scale levels in the high resolution image and from coarser scale levels in the coarse resolution image. The repeatability score is also influenced by rotation and illumination changes as well as the noise introduced by the camera. The results for large scale changes are 10% better than for the second best detector based on the Hessian matrix. However, the number of correct points is smaller for the Hessian detector. The repeatability of non adapted Harris detector is acceptable only for very small scale changes i.e. up to factor of 1.4. As we could expect,

Laplacian and DoG give similar results. Slightly better results for the LoG are caused by the noise and imprecision introduced by sampling of pyramid levels. The performance of scale invariant detectors is better compared to the Harris-Affine approach, but these detectors are adapted to the uniform scaling of the test images, whereas the affine detector can handle more complex image transformations.

4.3.2 Affine invariant detectors

In this section we present the evaluation results for the Harris-Affine, Harris-Laplace detector and the approach proposed by Schaffalitzky and Zisserman [104]. The last one is referred to as Harris-AffineRegions and it is in fact Harris-Laplace approach with an iterative procedure, which is applied to compensate for affine deformations of point neighborhoods. The location and scale of the point remain fixed during the iterations. We

0° 40° 50° 60° 70°

FIG. 4.10: *Images of one test sequence with perspective deformations. The corresponding viewpoint angles are indicated below the images.*

extended the evaluation criterion proposed for the scale change problem to the affine case. Similarly to the experiments carried out in the previous section the error in image surface ϵ_S covered by point neighborhoods is less than 20%:

$$\epsilon_S = 1 - \frac{\mu_a \cap (A^T \mu_b A)}{(\mu_a \cup A^T \mu_b A)} < 0.2 \tag{4.7}$$

where μ_a and μ_b are the elliptic regions defined by $x^T \mu x = 1$. The union of the regions is $(\mu_a \cup (A^T \mu_b A))$ and $(\mu_a \cap (A^T \mu_b A))$ is their intersection. A is a locally linearized homography H in point $\mathbf{x}_b$. We also neglect the possible 1.5 pixel translation error because the homography between real images is not perfect and the shift error has very little influence on ϵ_S. To simplify the computation we transform the points to obtain a unit circle of one of the regions and then we compute the intersection. The details can be found in annex A.2.

Figures 4.11 and 4.12 display average results for three real sequences of planar scenes (cf. figure 4.10, annex A.3). The viewpoint varied in horizontal direction between 0 and 70 degree. There are also some illumination and zoom changes between the images. The homography between images was estimated using manually selected corresponding points. Figure 4.11 displays the repeatability rate and figure 4.12 shows the localization and the intersection error for corresponding points. Corresponding points used for computing the errors are determined by the homography. In order to measure the accuracy of the localization we allow for points with up to 3 pixels location error relative to the homography,

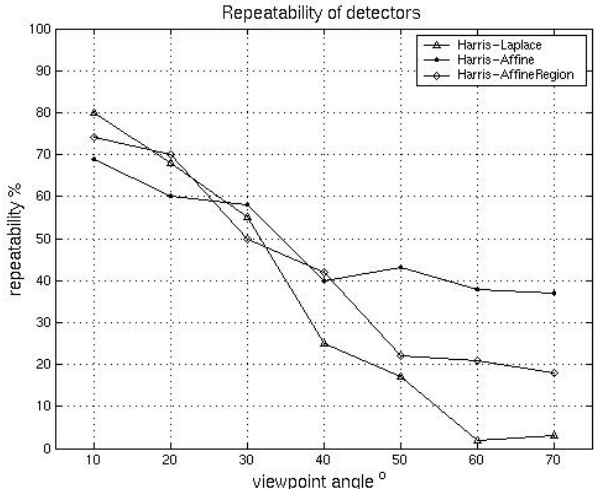


FIG. 4.11: *Repeatability of detectors: Harris-Affine - approach proposed in this paper, Harris-AffineRegions - Harris-Laplace detector with affine normalization of the point neighborhood, Harris-Laplace - multi-scale Harris detector with characteristic scale selection.*

and then estimate the average error. Similarly, to compare the error of region intersection, we allow for points with up to 40% maximal surface error and then compute the average error value. The real affine deformation is computed with a local approximation of the homography. We notice in figure 4.11 that our detector significantly improves the results in the case of strong affine deformations, that is for changes in the viewpoint angle of more than 40 degrees. The improvement is with respect to localization as well as region intersection (cf. figure 4.12). In the presence of weak affine distortions the Harris-Laplace approach provides better results. The affine adaptation does not improve the location and the point shape, because the scaling is almost the same in every direction. In this case the uniform Gaussian kernel is sufficiently well adapted. These results clearly show that the location of the maximum of the Harris measure and the extremum over scale are significantly influenced by an affine transformation.

4.4 Discussion

In this section we focussed on the problem of affine invariant detection of local features. We have proposed two novel approaches, which are the main contribution of this thesis. The approaches are based on a strong theoretical background, which was successfully explored by numerous researchers. The scale invariant detector was designed to handle the frequent problem of scale change between images taken from different distance or with different focal settings. This approach is based on two functions, the Harris and the Laplacian, both of which have been previously presented in literature, but separately. We have shown how to combine these two approaches to obtain the scale invariant interest point detector. It can reliably detect corresponding features in images related by a scale

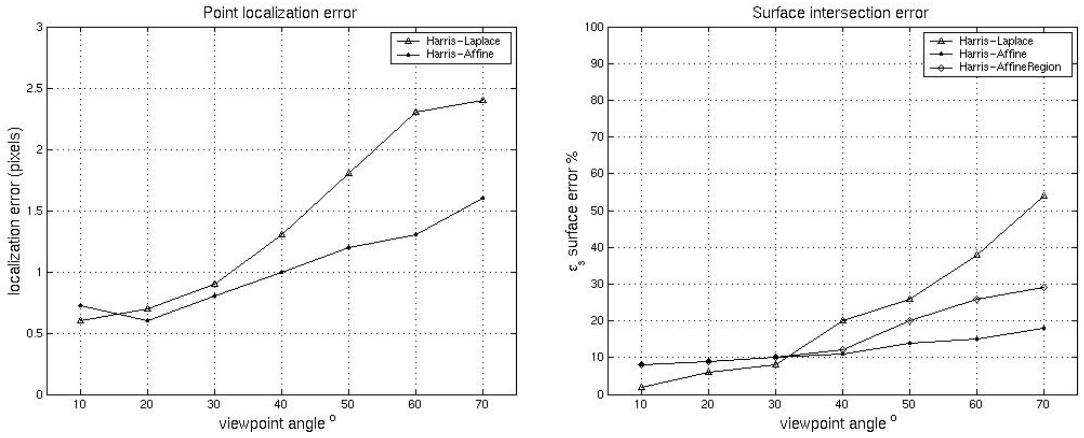


FIG. 4.12: *Detection error of corresponding points : (a) relative location (b) surface intersection* ϵ_S.

change up to a factor of 4. The location and the scale of interest points can be precisely detected by the iterative algorithm initialized with multi-scale Harris points. This detector can be used in the case of significant scale changes including weak affine deformations.

A general solution to affine photometric and geometric transformations including significant scale change was proposed in this chapter. The previous solutions handle the affine problem only partially and some of the important parameters are not estimated in an affine invariant way. Our algorithm simultaneously adapts the location, the scale and the shape of the neighborhood to obtain the affine covariant points. None of the existing methods simultaneously solves all of these problems in the feature extraction process. We believe that this approach will contribute in the improvement of matching and recognition.

The experimental evaluation carried out at the end of this chapter shows the excellent repeatability and accuracy of our approach. To obtain the representative results we have established the evaluation criteria and created a database of real image sequences with ground-truth. The gradual natural transformations between successive images enable one to verify, average and compare the detection results as well as evaluate the limits of different approaches.

The scale invariant detector gives better results than its affine extension in the case of uniform scale changes, but it fails when the scale change is significantly different in each direction. Therefore we propose using the Harris-Affine detector for general use and the Harris-Laplace detector, only when weak affine deformations are considered.

Chapter 5

Local image description

THE description of local image patterns is the next important step in the process of image recognition, after the extraction of features. The objective is to obtain a compact and complete feature description that enables a similarity measure to be applied. It is necessary for comparing similar quantities in images. The descriptors should capture the information about the shape and the texture of a local structure. The information content and the invariance are two important properties of local descriptors. These properties determine the distinctiveness and the robustness of the descriptors, and are reciprocally dependent. In general, the more the description is invariant the less information it conveys. The information content is the quantity of information conveyed by a descriptor. The maximum information content is determined by the detection function, which extracts points at some specific signal changes. Some of this information is usually discarded, while computing the descriptors. The local character that is the small size of features also limits the quantity of information.

There are numerous description techniques that can be used to compute the representation of a local feature. In section 5.1, we present the descriptors applied for local features, which recently appeared in literature in the context of scale and affine invariance. In section 5.2, we focus on the descriptors based on the derivatives computed in a point. We present the principal advantages and drawbacks of the differential invariants. Despite their sensitivity to different types of noise, their simplicity and accuracy in signal representation make them an appropriate descriptor for comparative evaluations of matching and recognition algorithms. Therefore, in section 5.3 we use the differential descriptors to compare the information content of interest points extracted with the scale and the affine invariant detectors.

5.1 State of the art

This section presents the descriptors, which have been proposed in literature to represent the scale and the affine invariant features. These descriptors represent geometric and photometric quantities, which determine the distinctive character of a local structure. The simplest descriptor is the image pattern itself. The cross-correlation measure can be applied to compute the similarity score between the regions. However, high dimensionality of such descriptors is prohibitive for recognition. In the context of database indexing, compact and low dimensional descriptors are required; otherwise the volume of the database is difficult to handle, while searching for similar descriptors. Therefore, this technique is usually used for finding point-to-point correspondences between two images. A different representation was introduced by Johnson and Hebert [54] and further developed in [55] in the context of object recognition in 3D scenes. Instead of using directly the pixel values, the object representation relies on descriptive images associated with points on the surface of the object. The images, called *spin-images*, are generated using the positions of pixels, defined by two parameters, relative to the interest point. The accumulation of these parameters for many points on the surface of the object results in an image for each interest point. The spin-image is a description of the shape of an object since it is the projection of the relative positions of 3D points into a 2D space. These images are invariant to rigid transformations of points. The description is also robust to partial occlusions due to the accumulated distribution of points. Cross-correlation measure can be used to compute the similarity between two spin images.

Lowe [73] (cf. section 4.1.1) proposed a descriptor based on the response properties of complex neurons in the visual cortex. To obtain the robustness to geometric distortions, a point neighborhood is represented with multiple images, which are, in fact, orientation planes representing each of a number of orientations. Each image contains only the gradients corresponding to one orientation with linear interpolation used for intermediate orientation. The images are smoothed and re-sampled to allow for larger shift in gradient localization. The dimension of the description vector is 160, which gives a possibility of rich and discriminative description. This description provides for some robustness against localization error and affine deformations, but the description vectors of dimension 160 require efficient searching algorithm in large databases.

Generalized color moments have been introduced by Mindru et al. [86] to describe the multi-spectral nature of the data:

$$M_{pq}^{abc} = \int_{\Omega} \int x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c dxdy$$

with order $p+q$ and degree $a+b+c$. These moments are independent and can be easily computed for any order and degree. The moments characterize the shape, the intensity and the color distribution in a limited region Ω. Usually invariants up to second order and first degree are used, which provide 18 descriptor components. Recently a new version of this descriptor appeared [85], which is invariant to affine illumination changes in RGB color space. Tuytelaars and van Gool [127, 128] used these descriptors to represent the affine invariant regions (cf. section 4.1.1).

Tell and Carlsson [118] and Matas et al. [81] proposed a similar approach to the affine problem. They developed an affine invariant descriptor for point pairs. The method consists in computing a description of the intensity profile along a line connecting two points. The points are detected with the classical Harris detector. The description is not robust unless the two points lie on the same planar surface. We have also seen in section 4.3 that the standard Harris detector is not robust to the affine transformation, as a result the localization of points changes within local neighborhood. If two points are separated by a discontinuity in depth, the profile changes when viewed from different angles. Furthermore, a 1-D profile does not carry much information compared to a 2-D pattern. A robust extension of this approach was recently proposed in [119]. Spatial relations between points are coded by a string, which describes the order of interest points in a semi local neighborhood. These relations are used to verify the correctness of matched points.

An approach, which has not yet been used in the context of scale or affine invariant features, but is interesting for its robustness, was developed by Zabih and Woodfill [136]. It relies on local transforms based on non-parametric measures, designed to tolerate partial occlusions. Non-parametric statistics [64] use the information about ordering and reciprocal relations between the data, rather than the data values themselves. Two local non-parametric transforms were introduced. The *rank transform* is a non-parametric measure of local intensity, and the *census transform*, is a non-parametric summary of local spatial structure. All non-parametric transforms depend on the comparison of the intensity of a point with the intensities of neighboring points. The *rank* transform is defined as a number of pixels in the region, whose intensity is less than the intensity of the center pixel. The *census* transform maps from the local neighborhood of a point to a bit string representing the set of neighboring pixels, whose intensity is inferior to that of the point. A small region is described by ordered binary relations between pairs of pixels. A robust description is represented by a distribution of numbers resulting from the transforms. It is, however, difficult to obtain the rotation invariance and the size of the descriptor vector is large. A similar approach to texture description was developed by Ojala [93]. An image region must be normalized to compensate for scale or affine transformations, before we compute such descriptor. Some promising results with non-parametric descriptors have been presented by Picard [96]. A rotation invariant version was proposed in this work. It would be, however, valuable to carry out a comparative evaluation of this descriptor on a large set of images.

There are many techniques based on a frequency content in an image. The image content can be decomposed into the basis frequencies by the Fourier transform, but the spatial relations between points are then lost as the Fourier transform is localized in frequency and not in space. Moreover, the basis functions are infinite, therefore difficult to adapt to the local approach. There is also the problem of artifacts appearing on the boundaries of a transformed local image pattern. To find the localization in space we can use the Gabor transform [39]. In this transform the signal is windowed with the Gaussian function, and then decomposed on the basis frequencies. It provides us with the information localized in space and in frequency. However, these are reciprocally dependent, and the precision limitation is the same as the Heisenberg inequality. The main problem with the Gabor filters is their dimensionality. To capture small changes in frequency and orientation,

a large number of filters is required. Therefore, much research is concentrated on reducing the number of results by combining or coding the filter responses [53]. The Gabor filters are frequently explored in the context of texture classification [9, 13, 28, 106, 134]. A set of Gabor filters for texture description is detailed in [79]. The wavelet transform provides another possibility to decompose the signal into basis frequencies. This transform and its properties are presented in section 7.2.1.

There is also a family of descriptors based on Gaussian derivatives. We analyze these descriptors in detail in the next section.

5.2 Differential description

In this section we present and analyze the differential descriptors. Section 5.2.1 presents different variants of descriptors based on local derivatives. We propose a method for estimating a dominant orientation in a local neighborhood of a point, which can be used to obtain the invariance to rotation (cf. section 5.2.2). Next, in section 5.2.3 we analyze the influence of noise on derivative responses. We discuss the similarity measure (cf. section 5.2.4) usually applied for this type of descriptors and the correlation between the descriptor components (cf. section 5.2.5).

5.2.1 Differential invariants

This description technique is used in our approach and it arises from the Taylor expansion, which locally approximates a differentiable function f. The function is approximated up to Nth order by:

$$f(x_0+x,y_0+y) = f(x_0,y_0) + x\frac{\partial}{\partial x}f(x_0,y_0) + y\frac{\partial}{\partial y}f(x_0,y_0) + \dots + \sum_{p=1}^{N} x^p y^{N-p} \frac{\partial^N}{\partial x^p \partial y^{N-p}} f(x_0,y_0) + \mathcal{O}(x^N, y^N)$$

This expansion makes use of a set of derivatives to describe a small image region. A stable estimation of local derivatives can be obtained by applying the Gaussian operators. The responses of Gaussian derivatives form a compact description. The vectors are of low dimension. Therefore, they are easy to manipulate in a large database and the computation of the similarity between two descriptors is fast. The invariance to geometric and photometric transformations can be obtained in different ways. However, in practice these descriptors suffer from important drawbacks, which we emphasize in the next sections of this chapter.

A set of derivatives computed at a point represents the geometric and photometric properties of the neighborhood of the point. The vector of such derivatives was named by Koenderink [60] *Local Jet*. For a given scale factor σ, *Local Jet* is defined by:

$$J^N[I](\mathbf{x},\sigma) = \{L_{i_1 \dots i_n}\mathbf{x},\sigma) | (\mathbf{x},\sigma) \in I \times R^+, n = 0,\dots,N\}$$

The scale parameter σ enables the descriptor for the scale to be computed, at which the interest point is detected. To obtain a complete local description we decompose the signal into a set of derivatives up to 4th order. Derivatives are computed on image patches centered in interest points. The masks applied for computing the derivatives are presented in figure 3.5. The local derivatives are computed in the directions of Cartesian image coordinates. One can assign to each feature a canonical orientation so that the descriptors are invariant to rotation (cf. section 5.2.2). Freeman and Adelson [38] and later on Perona [95] showed how to analytically compute the derivatives in a particular direction given the components of *Local Jet*:

$$\begin{aligned}
L'(\theta) &= L_x \cos(\theta) + L_y \sin(\theta) \\
L''(\theta) &= L_{xx} \cos^2(\theta) + 2L_{xy} \cos(\theta)\sin(\theta) + L_{yy} \sin^2(\theta) \\
L'''(\theta) &= L_{xxx} \cos^3(\theta) + 3L_{xxy} \cos^2(\theta)\sin(\theta) + L_{xyy} \cos(\theta)\sin^2(\theta) + \\
&\quad + L_{yyy} \sin^3(\theta) \\
L''''(\theta) &= L_{xxxx} \cos^4(\theta) + 4L_{xxxy} \cos^3(\theta)\sin(\theta) + 6L_{xxyy} \cos^2(\theta)\sin^2(\theta) + \\
&\quad + 4L_{xyyy} \cos(\theta)\sin^3(\theta) + L_{yyyy} \sin^4(\theta)
\end{aligned} \tag{5.1}$$

To represent a derivative of nth order, we compute $(n+1)$ directional derivatives oriented in $\theta_{n,i}$, $i = 0 \ldots n$ directions. The directions are $\theta_{n,i} = i\pi/(n+1) + \theta_g$, where θ_g is an orientation related to the image structure.

Our illumination change model permits affine transformation of pixel intensities $(aI(\mathbf{x}) + b)$. The translation factor b is eliminated by the differentiation operation. Invariance to the linear scaling by factor a is obtained by dividing the higher order derivatives by the first derivative:

$$\nu[0, \ldots, 12] = [\frac{L''(\theta)}{L'(\theta)}, \ldots, \frac{L'''(\theta)}{L'(\theta)}, \ldots, \frac{L''''(\theta)}{L'(\theta)}] \tag{5.2}$$

It is straightforward to show that these expressions are invariant to affine intensity changes:

$$\frac{\frac{\partial^2}{\partial x^2}(aI(\mathbf{x}) + b)}{\frac{\partial}{\partial x}(aI(\mathbf{x}) + b)} = \frac{a\frac{\partial^2}{\partial x^2}I(\mathbf{x})}{a\frac{\partial}{\partial x}I(\mathbf{x})} = \frac{\frac{\partial^2}{\partial x^2}I(\mathbf{x})}{\frac{\partial}{\partial x}I(\mathbf{x})} \tag{5.3}$$

Using the derivatives up to 4th order, we obtain descriptors of dimension 12.

The differential invariants combining the components of *Local Jet* (cf. equation 5.4) have been introduced by Koenderink [60] and ter Haar Romeny [120]. These descriptors are invariant to rotation. They can also be invariant to affine illumination changes by eliminating the first two components of the vector 5.4 and dividing the other ones by the appropriate power of the second component, in a similar way as was done for the steerable filters.

$$\begin{bmatrix}
L \\
L_xL_x + L_yL_y \\
L_{xx} + L_{yy} \\
L_{xx}L_xL_x + 2L_{xy}L_xL_y + L_{yy}L_yL_y \\
L_{xx}L_{xx} + 2L_{xy}L_{xy} + L_{yy}L_{yy} \\
L_{xxx}L_yL_yL_y + 3L_{xyy}L_xL_xL_y - 3L_{xxy}L_xL_yL_y - L_{yyy}L_xL_xL_x \\
L_{xxx}L_xL_yL_y + L_{xxy}(-2L_xL_xL_y + L_yL_yL_y) + L_{xyy}(-2L_xL_yL_y + L_xL_xL_x) + L_{yyy}L_xL_xL_y \\
L_{xxy}(-L_xL_xL_x + 2L_xL_yL_y) + L_{xyy}(-2L_xL_xL_y + L_yL_yL_y) - L_{yyy}L_xL_yL_y + L_{xxx}L_xL_xL_y \\
L_{xxx}L_xL_xL_x + 3L_{xxy}L_xL_xL_y + 3L_{xyy}L_xL_yL_y + L_{yyy}L_yL_yL_y
\end{bmatrix} \tag{5.4}$$

Baumberg [7] and later on Schaffalitzky and Zisserman [104] proposed to use a filter bank derived from the family of Gaussian operators G:

$$K_{mn}(x,y) = (x+iy)^m(x-iy)^n G(x,y)$$

Under a rotation by angle θ, the two complex coordinates $z = x + iy$ and $\hat{z} = x + iy$ transform as $z \to e^{i\theta}z$ and $\hat{z} \to e^{-i\theta}\hat{z}$, where x and y are Cartesian coordinates. The filters K_{mn} are then multiplied by $e^{i(m-n)\theta}$. These filters differ from the Gaussian derivatives by a linear change of coordinates in filter response space. The magnitude of response is not influenced by the transformations, but only the phase changes. They use 16 filters defined by combinations of m and n. This approach avoids the problems related to the estimation of a dominant orientation of a local feature, but simultaneously limits the number of possible filters in comparison with directional derivatives. In theory, the filters are orthonormal, therefore the Euclidean distance can be used to compute the similarity score.

5.2.2 Dominant orientation estimation

The invariance to rotation can be obtained either by computing rotation invariant descriptors or by normalizing the region with respect to rotation and then computing the description. Rotation invariant photometric measures were proposed in [38, 60, 120]. A rotation invariant description can also be obtained with circularly symmetric operators, as for example the Laplacian, but the number of such operators, which are useful in practice, is limited. A stable estimation of one dominant orientation in a local neighborhood is required, in order to normalize the neighborhood to rotation. If the estimation is incorrect, the computed description is useless, as it is not rotation invariant and cannot be correctly matched. Furthermore, if there are many useless descriptors the probability of a false match for the remaining correct descriptors is increased. This technique is less robust against noise and arbitrary image transformations but makes it possible to apply descriptors which are not invariant to rotation. The estimation of dominant orientations is often based on the phase of the gradient, computed in the interest point.

Lowe [73] proposed a histogram based approach. The local gradient orientations within a small region around a point are accumulated in a histogram. Each histogram bin corresponds to one of 36 possible orientations uniformly distributed in the full range of 360 degrees. The influence of a gradient phase in a point on a histogram bin is weighted with the gradient value in that point. A strong gradient is more robust against noise thus the phase of this gradient is more stable. On the other hand there is some risk of incorrect estimation, when a histogram bin is significantly increased by a point relatively distant from the feature center, but with a very strong gradient value. Therefore, Lowe proposed to weight also the gradient value with a Gassian window. The orientation corresponding to the largest bin in the histogram is selected to compensate for the rotation.

To obtain a stable estimation of the dominant direction in point $\mathbf{x}_0$, we propose to use the average orientation computed in the point neighborhood. The phase deviations in neighboring points are weighted by the Gaussian window centered in the point $\mathbf{x}_0$. Thus, we decrease the influence of points further from the center of the region. Next, the average

gradient deviation is computed in the nearest neighborhood of the interest point. Finally, we correct the gradient phase $\theta(\mathbf{x}_0) = \arctan(\frac{L_y(\mathbf{x}_0)}{L_x(\mathbf{x}_0)})$ with the average deviation:

$$\theta_g = \theta(\mathbf{x}_0) - \frac{1}{\sum_m g(\mathbf{x}_m, \sigma_I/3)} \sum_m g(\mathbf{x}_m, \sigma_I/3)(\theta(\mathbf{x}_0) - \theta(\mathbf{x}_m))$$

The size of the Gaussian window is 3 times smaller than the size of the feature. This factor was experimentally selected and gave the best estimation results. The experimental results show that this technique is more robust to the localization error, affine deformations and contrast changes than simple $\theta_g = \theta(\mathbf{x}_0)$ computed in the interest point.

Figure 5.1 shows the results of comparison between our approach, the histogram based approach and the orientation computed exactly in the point $\mathbf{x}_0$. To carry out this test we used several real image sequences with the corresponding points established by a homography matrix. The homography was also used to recover the real rotation angle between images. Figure 5.1(a) shows an example of an interest point detected near a corner. The

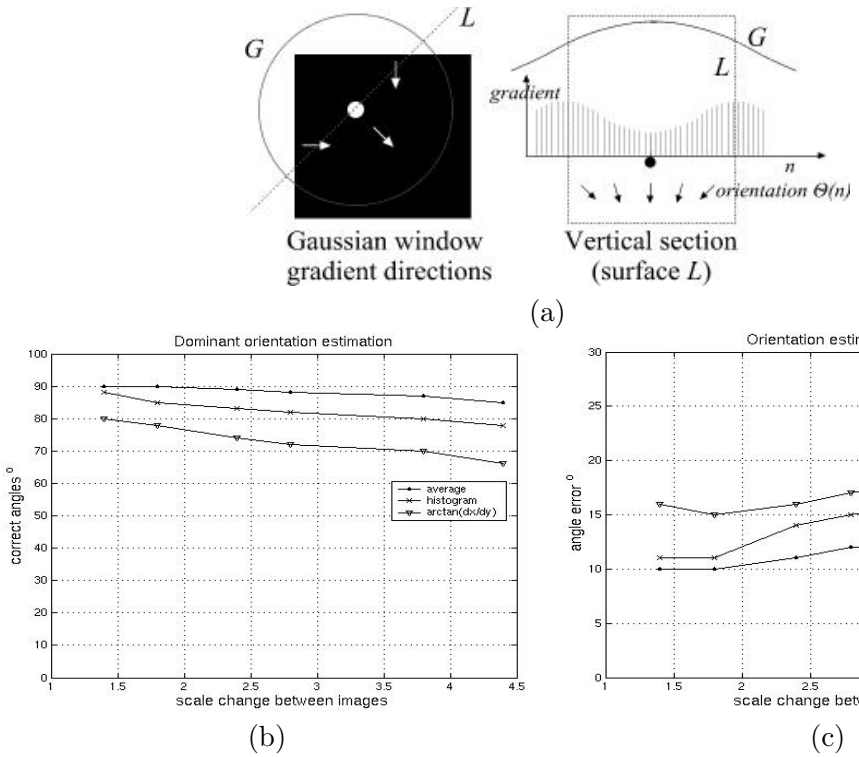


FIG. 5.1: *(a) Orientations in a point neighborhood. (b) Percentage of points with correctly estimated dominant orientation. (c) Mean error of the orientation angle for corresponding points.*

gradient amplitude and phase along the section L is shown on the right of the figure. We can see that the gradient value is higher on the edges forming the corner. The phase of the gradient is different for the edges and for the interest point. Computing an average

angle is more stable than selecting one of the dominant phases. Figure 5.1(b) shows the percentage of points for which the angle is correctly estimated. We can see that the method based on the average angle gives a higher percentage of correct estimations, although at least 10% of features is lost because of incorrect orientation. Figure 5.1(c) shows the average angle error. For our method, this error does not exceed 15 degrees. The gradient orientation is less distinctive for structures different to corners. Blobs or multi-junction features have more than one dominant orientation. Therefore, the estimation is more often incorrect for these features than for the corners. We can see that the method based on the average gradient deviation is slightly more stable than other techniques. The main drawback of the histogram based approach is that when there are two different dominant orientations with comparable gradient values a noise will determine the higher peak in the histogram. A possible solution is to compute one descriptor for each orientation. The average orientation is less accurate but it is also less influenced by the noise.

5.2.3 Noise

A common drawback of the differential descriptors is that they are very sensitive to localization error and high frequency noise. The influence of these errors on the derivative response can be explained by mathematical equations. In the following we derive the relations between the error and the filter response. Consider a function $I(x)$ and a periodic noise $\xi_1 = \varepsilon \sin(\omega x)$.

$$\hat{I}(\mathbf{x}) = I(\mathbf{x}) + \varepsilon \sin(\omega \mathbf{x})$$

The functions $I(\mathbf{x})$ and $\hat{I}(\mathbf{x})$ are similar if the noise can be ignored that is, if its amplitude is small with respect to the amplitude of the signal. However, the first derivative $I_x(\mathbf{x})$ can be very different from $\hat{I}_x(\mathbf{x})$ if ω is sufficiently large ($\varepsilon \ll \omega$).

$$\hat{I}_x(\mathbf{x}) = I_x(\mathbf{x}) + \omega\varepsilon \cos(\omega \mathbf{x})$$

Notice that high frequency noise can significantly modify the derivatives. To reduce this effect we eliminate the high frequencies by computing the derivatives with Gaussian filters, which are in fact low-pass filters. The Gaussian derivatives also reduce the random noise by smoothing the signal.

The second source of noise is the localization error of points, in which the derivatives are computed. For a given interest point (x_0, y_0), the error for the first derivative can be computed using the Taylor expansion:

$$\begin{aligned} \xi_2 = L_x(x_0 + dx, y_0 + dy) - L_x(x_0, y_0) \approx \\ L_x(x_0, y_0) + dx L_{xx}(x_0, y_0) + dy L_{xy}(x_0, y_0) - L_x(x_0, y_0) = \\ dx L_{xx}(x_0, y_0) + dy L_{xy}(x_0, y_0) \end{aligned}$$

We can see that for the differential invariants the error ξ_2 is equal to the higher order derivatives multiplied by the displacement error. Note that the signal variations are strong in the interest points, thus the derivative values are significant. Given the points detected with a finite accuracy, we cannot reduce this error without any prior knowledge of the

exact location of the interest points. Therefore, in practice, this error mainly influences the point description. It increases the standard deviation of the *Local Jet* components and introduces the correlation between some of them (cf. section 5.2.5).

5.2.4 Distance measure

An algorithm for finding similar descriptors must rely on a similarity measure which enables a distance between two descriptors to be computed. The components of descriptors can vary in different range of values and can be correlated with each other. Therefore, a similarity measure should be adapted to the type of descriptor. There are numerous ways of measuring the similarity between features. One could apply the Euclidean measure L_2 or χ^2 test to compare two histogram distributions.

$$\mathbf{d}^2_{L_2} = ||\mathbf{v}_1 - \mathbf{v}_2||^2, \qquad \chi^2 = \sum_i \frac{(v_{1_i} - v_{2_i})^2}{(v_{1_i} + v_{2_i})^2} \tag{5.5}$$

The Euclidean distance can be used if the components of descriptors are not correlated and the variables change in the same range of values. We can apply different cross-correlation measures (SSD-sum of squared differences, ZNCC zero-mean cross-correlation) for descriptors using pixel values directly, although the cross correlation is difficult to use in the context of databases. A similarity measure must take into account different standard deviations and correlations of descriptor elements computed for one class of local structures. A class is a set of features representing the same image structure, but viewed in different conditions. The Mahalanobis distance is the appropriate measure for comparing two vectors of differential invariants.

$$d^2_M(\mathbf{v}_1, \mathbf{v}_2) = [\mathbf{v}_1 - \mathbf{v}_2]\Lambda^{-1}[\mathbf{v}_1 - \mathbf{v}_2], \qquad \Lambda^{-1} = P^T D^{-1} P = (D^{-\frac{1}{2}}P)^T D^{-\frac{1}{2}} P$$

The variance-covariance matrix Λ normalizes the distance using standard deviations and correlations of vector components. Computing this distance for all descriptors in a large database is very time consuming. Therefore, given the Λ matrix we can normalize all the vectors in the database with:

$$\mathbf{v}_{norm} = D^{-\frac{1}{2}} P \mathbf{v} \tag{5.6}$$

After this, we can apply the fast Euclidean measure every time we look in the database for similar descriptors.

$$d^2_M(\mathbf{v}_1, \mathbf{v}_2) = D^{-\frac{1}{2}} P(\mathbf{v}_1 - \mathbf{v}_2), \qquad \mathbf{v}_{norm} = D^{-\frac{1}{2}} P \mathbf{v}$$

One of the properties of the Mahalanobis distance is that it assumes Gaussian distribution of each of the descriptor elements computed for one class of features. The square of Mahalanobis distance follows the χ^2 distribution. Mathematical tables of χ^2 function enable the selection of the distance, within which we find a chosen percentage of descriptors of one class of features. However, the Gaussian function roughly approximates the distribution of differential invariants computed for a real data. The descriptors are more dispersed in the multi-dimensional space because of noise, the variations in photometry,

the inaccuracy of interest point locations and so forth. The distance between descriptors of the same class roughly follows the χ^2 distribution and the theoretical threshold is not adequate. Therefore, it must be determined empirically. Similarly, the variance-covariance matrix cannot be estimated by an analytical analysis. It is impossible to create a general model for all possible characteristic image structures, because a feature can take any form of a signal change. Therefore, an analytical simulation of descriptor behavior is possible only for simple, not textured structures such as perfect corners, edges, or Gaussian blobs. The same holds for the noise model, which is usually unpredictable. Consequently, the covariance matrix also has to be estimated on the real data.

5.2.5 Variance and covariance of differential invariants

The distance measure requires the estimation of the covariance matrix Λ which incorporates the variations $\sigma^2_{v_a}$ and correlations $\sigma_{v_a v_b}$ of descriptor components. In the case of discrete data these components can be approximated by:

$$\sigma^2_{v_a} = \frac{1}{n}\sum_i^n (E(v_a) - v_{a_i})^2, \quad \sigma_{v_a v_b} = \frac{1}{n}\sum_i^n (E(v_a) - v_{a_i})(E(v_b) - v_{b_i})) \tag{5.7}$$

where $E(v_a) = \frac{1}{n}\sum_i^n (v_{a_i})$ is a mean value of the components v_a. To recover the correlation between the elements we have to normalize the covariance by the variance. The normalized correlation factor of the components v_a and v_b is defined by:

$$\rho_{v_a v_b} = \frac{\sigma_{v_a v_b}}{\sqrt{\sigma^2_{v_a}\sigma^2_{v_b}}} \tag{5.8}$$

The correlation factor has several useful properties:

1. $-1 \leq \rho_{v_a v_b} \leq 1$
2. if v_a is linearly dependent on v_b:

$$v_a = \eta v_b + \epsilon \Rightarrow \begin{cases} \rho_{v_a v_b} = +1 & si \quad \eta > 0 \\ \rho_{v_a v_b} = -1 & si \quad \eta < 0 \end{cases}$$

3. if v_a and v_b are independent $\rho_{v_a v_b} = 0$,
4. if $\rho_{v_a v_b} = 0$ the variables are not correlated but they can be dependent.

We obtain a symmetric correlation matrix using equations 5.7:

$$\rho = \begin{pmatrix} 1 & \cdots & \rho_{v_n v_0} \\ \vdots & \ddots & \vdots \\ \rho_{v_0 v_n} & \cdots & 1 \end{pmatrix}$$

Ideally, the components of a descriptor vector are independent. Thus, the description is compact and not redundant. To identify the source of correlation we first detected the

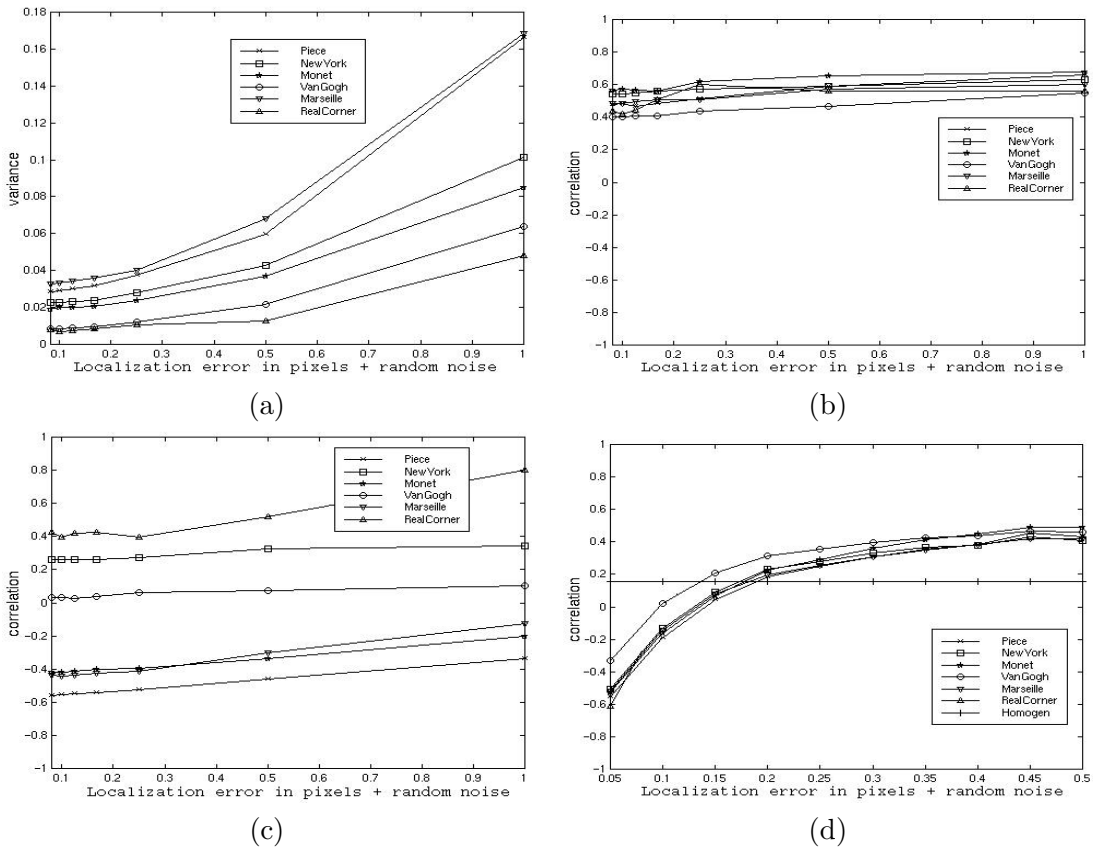


FIG. 5.2: *(a) The variance of the 2nd element of the vector represented by equation 5.4 with respect to localization error and noise, (b) Example of a correlation issued from mathematical relations (between 2rd and 3th element), (c) Example of correlation dependent on the texture (between 2nd and 5th element), (d) Example of correlation dependent on the shift error of the interest point (between 4th and 8th element).*

interest points in images containing different textures and than established the point-to-point correspondences using the homography between images. Next, we introduced the high frequency noise and the localization error to the interest points (cf. figure 5.2). The descriptors computed on these points are used to estimate the variance and the covariance. The results are presented in figure 5.2. Each curve represents a different image therefore a different texture. The essential information is represented by the shape of curves. The increasing values means that the variance or covariance is dependent on the localization error or the noise introduced to the interest points. Figure 5.2(a) shows that the instability is due mainly to the localization accuracy of the interest points. The description is less influenced by the noise, or by the illumination changes different to the affine model. There are, in fact, three types of correlation. The first one is the mathematical relation between the invariants. A constant, but different from zero value, stands for the correlation introduced by the differential equations. This type of correlation is presented in figure 5.2(b). It is independent of noise and geometric and photometric changes. It may arise from the fact that the differential expressions combine the same components of the *Local Jet* (cf. equation 5.4). Different and slightly increasing values presented in figure 5.2(c) signify that the correlation depends on the texture in images and on the localization error of interest points. Many texture motifs are repeated in the images, thus the computed invariants are similar. The curves shown in figure 5.2(d) are the same for all types of textures, but change with respect to the localization error.

The results presented in figure 5.2 confirm the assumption that the correlation must be estimated on the real data. The correlation can be used to transform the coordinate system of descriptors. It is, however, more prudent to set the correlation between some components to zero, given that it changes with respect to the texture in images or to the localization error. If the correlation depends only on the mathematical relations it can be used to generate the covariance for different values of standard deviation. This gives a certain degree of freedom for changing the size of a class of descriptors. It allows for a different size of classes in the descriptor space. Given the variance and the correlation between two components we can compute the covariance:

$$\sigma_{v_a v_b} = \rho_{v_a v_b} \sqrt{\sigma^2_{v_a} \sigma^2_{v_b}} \tag{5.9}$$

Thus, we can associate a different covariance matrix to each class of descriptors and use them to compute the distance between a new descriptor and the class. The distance is then more accurate, as the covariance matrix is adapted to the class.

For all other experiments described in this manuscript the covariance matrix Λ is estimated statistically over a large set of real image samples. It incorporates signal noise, variations in photometry, inaccuracy of interest point location, and so forth. To estimate this matrix we used the images containing planar scenes with homography computed on manually selected points. Next, we detected the invariant local features and determined the point-to-point correspondences with the homography matrix. The descriptors of the points were finally used to compute the variance and covariance of the vector components. Next, we removed the unstable correlation identified by the analysis presented in this section. The matching tests for different images gave slightly, but systematically better results.

5.3 Entropy of descriptor space

One of the most important properties of an image feature is the quantity of information that it conveys. The distinctiveness of interest point descriptors depends on this information and can be measured by a criterion introduced in [109]. In the following we describe this criterion and present the experimental results.

5.3.1 Entropy criterion

Ideally, we want to have a very compact distribution of descriptors within the same class, and a large distance between different classes. The dispersion of descriptors can be measured by its entropy. We can interpret the descriptor space as a space of all possible messages and a feature descriptor as a single message. The entropy is the average information content per message:

$$E(\mathcal{P}) = -\sum_i p_i \ln(p_i)$$

where $\mathcal{P}$ is a partition of the space in cells, and p_i is a probability that a descriptor occurs in the cell i. We assume that one cell represents one distinctive message, therefore it contains only descriptors of similar image structures in terms of the Mahalanobis distance. In practice the probability p_i is measured by the number of descriptors in the cell to the overall number of points in the descriptor space. The higher the cell probabilities the lower the entropy, therefore the distinctiveness of the descriptors is lower. If all the descriptors are located in one cell the entropy is 0, and the descriptors convey no salient information. The entropy is high if the descriptors are dispersed uniformly in the space. Note that we can obtain different entropy for different partitions, which are determined by the cell size. The cell size should not be larger than one class of descriptors, otherwise the dispersion is not reliably measured. To compare the information content of points extracted with different detectors we apply the same descriptor and the same uniform partition to all points. In order to assign the points to the cells with the Euclidean distance, the descriptor vectors are normalized with the covariance matrix (cf. equation 5.6).

5.3.2 Results for information content

In table 5.1 we show the entropy of interest points taken from different detectors. We used about 100000 points for computing the entropy for each of the detectors. Each dimension is partitioned into the same number of cells of the same size. The normalizing covariance matrices are estimated separately for the steerable filters and differential invariants. The first row indicates the detectors used to extract the interest points. The second row shows the entropy results for the differential invariants. In the third and the fourth row we can see the entropy of descriptors computed with the steerable filters. The steerable filters obtain better entropy in the descriptor space than the differential invariants. This might be due to the correlation of the invariants. We notice that the Harris-Laplace points convey more information than the Harris multi-scale points. The probability of false match

is $e^{-(8-10.2)} \simeq 9$ times higher for the Harris multi-scale than for the Harris-Laplace. These results justify the name *characteristic* for the points with scale selected by the Laplacian. The Laplacian combined with the Harris measure enable the neighborhood to be captured, where the signal change, measured with the first and the second derivative, is larger and the information content is therefore higher. We also estimated the entropy of descriptors computed for the localization and the scale of Harris-Affine points, but the point neighborhood was not affine normalized. The entropy for these points is higher than for the points with affine normalized regions. It shows how the affine normalization reduces the information content. We could expect these results because obtaining invariance to any transformation consists in removing the information about this transformation. The entropy is computed for descriptors based on derivatives up to 3rd order. The higher entropy for invariants computed up to 4th order shows that higher order derivatives can be useful despite their sensitivity to noise. However, one should be careful applying higher order derivatives as their usefulness is significantly lower for the Gaussian kernels of a small size. The Laplacian function detects mainly the blob-like structures, therefore the entropy is slightly lower than for Harris-Laplace points.

derivatives up to order	Harris-multi-scale	Harris-Laplace	Harris-Affine	Harris-Affine (loc.,scale)	LoG 3D extrema
3rd (7 invariants)	5.9	7	6.2	6.5	6.4
3rd (7 components)	6.9	8	7	7.6	7.3
4th (12 components)	8	10.2	8.8	9.6	9

TAB. 5.1: *Entropy of interest points extracted by different detectors. The higher the entropy the more distinctive the interest points. The top row indicates the interest point detector. The second row shows the entropy for differential invariants. The third and fourth rows show the entropy for steerable filters*

5.4 Discussion

In this chapter we have focused on the differential descriptors. These descriptors are very compact, easy to implement, fast to compute and simultaneously they can represent a local image structure well. It is also straightforward to apply an appropriate metric to compute a similarity distance. Unfortunately, the differential descriptors suffer from many drawbacks.

In section 5.2.3 we analyze the noise that mainly influences these descriptors. We have studied the components of the variance-covariance matrix in order to identify the source of instability of descriptors. A relatively small localization error can significantly influence the derivative value and consequently the descriptor components. The differential invariants are also very sensitive to high frequency noise. The descriptors based on Gaussian kernels represent essentially the information in the center of the feature, which is due to the nature of the Gaussian function. The experimental results showed different types of correlation introduced by mathematical relations, texture, inaccuracy in location of interest points

and image transformations. Therefore, the estimation of the covariance matrix should be done empirically on real data.

The differential invariants can be applied for description if a high accuracy of feature localization is expected. These descriptors are suitable for experimental analysis and give reliable results in relative comparisons as the results are equally influenced by different kinds of noise.

We have proposed and evaluated a stable method for estimating the dominant orientation in a local neighborhood. The orientation is estimated independently of the description. The orientation is related to the local structure, therefore it can be used to rotate the image pattern before computing the description. Hence, any descriptor computed on rotated patches is invariant to rotation.

The same operation can be done to obtain the invariance to affine illumination changes. We can either compute invariant descriptors or locally normalize the image patch and then compute the description. It is not obvious as to which of these two methods is more robust, therefore a comparative evaluation on a large set of real images would be valuable.

We have carried out the comparative evaluation and presented the results for the information content of the scale and the affine invariant features. The Harris-Laplace detector obtained the best results in this comparison. All Harris-Affine features are isotropic, since this detector removes the information about the affine deformation of the patterns. Therefore, the entropy for these features is lower. The entropy is higher for steerable filters than for differential invariants. This is due to the correlation between vector components and the multiplication of noise, which is more exposed by the differentiation.

The evaluation of differential descriptors, which is presented in this chapter, showed the necessity for applying a more efficient feature description. One notices that we do not use the information represented by color. A descriptor using color can certainly better represent a local image structure. However, color information is very sensitive to illumination changes. The information represented by color can be significantly changed by shadows, light color, the direction of light source and so forth. Therefore to obtain the invariance to these changes a more complex model than the simple affine one must be applied. The descriptors should also capture information about the texture within a point neighborhood. The frequency content can be extracted by Gabor filters or by the wavelet decomposition. It is also possible to use the non-parametric descriptor (cf. section 5.1), but the high dimensionality of these descriptors should be reduced to make it useful in practice. Several suggestions for further investigation are outlined in section 8.2.

Chapter 6

Matching and indexing

MATCHING and recognition are the most important applications for local features. Many reconstruction algorithms rely on matching results. A reliable and fast recognition approach is one of the most required solutions in the computer vision. Unlike global approaches the interest points can be invariant to change in the content of viewed scenes and occlusions. The local approach significantly improved and accelerated the recognition process.

In this chapter we present the experimental results for matching and image retrieval. In section 6.1 we present our matching and indexing method. We first present the steps of the algorithm, which are common for the considered applications. Each of the following sections presents one part of the algorithm and the essential parameters related to this part. Next, we explain the steps, which are different for the matching and for the indexing process. Section 6.2 presents the results for matching images with significant transformations. The performance of image retrieval approach is evaluated in section 6.3.

6.1 Matching and indexing algorithm

The detection of interest points is the first step in the matching and the recognition process and is described in sections 6.1.1 and 6.1.2. In the context of matching the points are extracted from two images. Next, we compute the description for each of the interest points (cf. 6.1.3). To establish the point-to-point correspondences we use the similarity measure, which is described in section 6.1.4. Finally, we apply a robust algorithm to estimate the transformation between the images. This algorithm is presented in section 6.1.5.

In the case of image retrieval we first create a database of descriptors computed for interest points detected in all images used for our experiments. Each entry in the database contains a descriptor with a pointer to the image, in which the descriptor was computed. Given a query image, we apply the algorithm described in section 6.1.6 to find the most

similar model in the database. In the following sections we discuss, in detail, each step of the algorithm.

6.1.1 Detection of scale covariant points

The detection starts with initial points extracted with multi-scale Harris detector. Next, we select the scale invariant point with the Laplacian measure (cf. equation 4.2), as described in section 4.1.2. The derivation and integration scales are related by: $\sigma_D = 0.6\sigma_I$. The points are extracted at 17 scale levels for $\sigma_I = 1.2^n$ with $n = 1, \ldots, 17$. The parameter α in the Harris function is set to 0.06 (cf. equation 4.1). The threshold for the Harris measure is set to 1000, and the threshold for the Laplacian measure is set to 10. We keep all the points, which are maxima with respect to the Harris and the Laplacian measure, even if they represent the same structure. The exact scale and location of each structure can be found with the iterative algorithm described in section 4.1.2. The points are represented by the coordinates (x, y) in the image, and the scale σ_I. Usually, there are a few hundred interest points per image representing an outdoor scene, which is almost ten times less than the number of multi-scale Harris points. The extraction of Harris-Laplace points in an image of size 800×600 at the Pentium II 500MHz takes one minute.

6.1.2 Detection of affine covariant points

The affine invariant detector is presented in section 4.2.2. The initial points are detected by multi-scale Harris detector. The scale levels and the threshold are the same as for the Harris-Laplace detector, and are given in the previous section. Next, for each point we apply the iterative algorithm. Without loss of generality we restrict the range of the parameters used to estimate the affine regions. This enables us to limit the search space and therefore to accelerate the procedure. The integration scale is searched among the values distributed exponentially $\sigma_I^{(k)} = s\sigma_I^{(k-1)}$, where s is computed with $s = 0.4 \cdot 1.2^n$ and $n = 0 \ldots 7$. The derivative scale is searched among the values $\sigma_D^{(k)} = k\sigma_I^{(k)}$, with $k = 0.4 \cdot 1.12^n$ and $n = 0 \ldots 5$. The termination criterion is set to $\epsilon_C = 0.05$ (cf. equation 4.4). The maximal stretching of the region is defined by $\epsilon_l = 6$ (cf. equation 4.5). We have applied a set of parameter values that enables a stable set of affine invariant points to be obtained in an acceptable computation time. The points are represented by the coordinates (x, y) in the image, the characteristic scale σ_I and the affine transformation matrix U (cf. equation 4.3). Usually, about 40% of initial points do not converge due to the lack of characteristic scales or to the large difference between the eigenvalues of the matrix U ($\epsilon_l > 6$, cf. equation 4.5). About 30% of the remaining points is selected by the similarity measure. Similar points are eliminated by comparing location, scale and second moment matrices as described in section 4.2.2. We can suppose that the features are more representative if they are present at a wide range of scales. These features are identified by several points, which have converged to the same structure. Finally, 20-30% of initial points provided by multi-scale Harris detector are used to represent an image. The extraction of Harris-Affine points in an image of size 800×600 with about 1000 initial points, takes a few minutes at the Pentium II 500MHz.

6.1.3 Description

A local feature is represented by a set of derivatives computed with the filters shown in figure 3.5, and re-oriented in the direction related to the local structure by the steerable filters (cf. equation 5.1). The dominant orientation is estimated with the method described in section 5.2.2. We use a vector of 12 elements, invariant to affine intensity changes (cf. equation 5.3) and rotation. In the case of affine invariant features the regions are normalized with estimated affine transformation, before computing the description. In the case of scale invariant points the descriptors are computed on the original image. These descriptors are used for matching and indexing the images.

6.1.4 Similarity measure

The similarity of descriptors is measured by the Mahalanobis distance. The invariant vectors are transformed with inverse covariance matrix as shown in section 5.2.4 (cf. equation 5.6). This permits the Euclidean distance to be used and significantly accelerates the computation. The covariance matrix and the similarity threshold are estimated on a large set of image samples. More details can be found in section 5.2.5. We apply the cross-correlation measure (SSD) for an additional verification. The points are then normalized with respect to the scale, the dominant orientation and the affine transformation. This enables the points, which have accidentally matched, due to similar differential descriptors to be rejected.

6.1.5 Robust matching

In general, we follow the classical matching algorithm. To robustly match two images, we first determine point-to-point correspondences using the similarity measure. We select for each descriptor in the first image the most similar descriptor in the second image. If the distance is below a threshold the match is potentially correct. The threshold is experimentally set at 15 to obtain approximately 50% of correct matches. A set of initial matches is obtained. In the second phase of verification we apply the cross-correlation measure, which rejects less significant matches. Finally, a robust estimation of the transformation between the two images based on RANdom SAmple Consensus (RANSAC) enables the selection of the inliers to the estimated transformation. The estimation results are reliable if more than 50% of matches are correct, although in practice the method correctly converges even with less matches. In our experiments the transformation is either a homography or a fundamental matrix. A model selection algorithm [56, 125] can be used to automatically decide which transformation is the most appropriate one.

6.1.6 Retrieval algorithm

In the following we outline the image retrieval algorithm. The recognition process consists in finding the most similar model represented by an image in a database. The database contains the normalized descriptors of image models. Given a query image we extract the interest points with one of the proposed methods and compute the normalized

differential description for the points. Next, for each of these points we look for similar descriptors in the database using the Euclidean distance. The distance threshold for similar points is set at 20. We can allow for more false matches than in the matching algorithm to retrieve more images with similar descriptors. The false images can be rejected by applying additional constraints [30, 99, 119]. A voting algorithm [113] is used to select the most similar images in the database. If the distance is less than the threshold, a vote is added for the corresponding model in the database, that is, the probability that the model is correct is increased. Note that a point cannot vote several times for the same database image, but the same image point can vote for different models. The model that obtained the highest number of votes is considered as the most similar to the query image. This makes retrieval robust to mismatches as well as to outliers.

The search for similar descriptors in the database is sequential. There are different ways of accelerating the search by changing the structure of the database [116, 129, 130, 132]. However, the problem of indexing high dimensional descriptors is not solved. In high dimensional databases the sequential search can be even faster than other techniques [4]. Although, our descriptor is of low dimension, the study presented in section 5.2 showed the necessity for a new more robust descriptor, probably of higher dimension. Moreover, the performance of the sequential search is sufficient for the evaluation of local features in the context of image recognition, which is the purpose of this part of the work.

6.2 Experimental results for matching

In this section we present the matching results for the approaches described in the previous section. Section 6.2.1 shows the results for the scale invariant approach applied to images with significant scale changes. In section 6.2.2 we apply the affine invariant method to images with strong perspective deformations.

6.2.1 Scale change

In the following, we apply the points detected with the Harris-Laplace method for matching images with significant scale changes. Figure 6.1 illustrates the consecutive steps of the matching algorithm. In this example two images are taken from the same viewpoint, but with a change in focal length and camera rotation. The top row shows the interest points. There are 190 and 213 points detected in the left and right images, respectively. These numbers are about equivalent to the number of points, which are usually detected with the standard Harris detector applied at the finest level of scale-space representation. Usually, there are about 10 times more points in the entire representation if the standard Harris detector is used. This clearly shows the selectivity of our method. If no scale peak selection had been used, more than 2000 points would be detected for 17 resolution levels. Column (b) shows the 58 matches obtained after the initial matching phase. Column (c) displays the 32 inliers to the estimated homography, all of which are correct. The estimated scale factor between the two images is 4.9 and the rotation angle is 19 degrees.

Figure 6.2 shows an example for a 3D scene where the fundamental matrix is robustly estimated. There are 180 and 176 detected points detected in the left and the right image.

The number of initial matches is 23 and there are 14 inliers to the estimated fundamental matrix, all of them correct. Note that the images are taken from different viewpoints and the transformation includes a scale change, an image rotation as well as a change in the viewing angle.

6.2.2 Significant viewpoint change

Figure 6.3 illustrates the results of the matching procedure for Harris-Affine interest points. In order to separate the detection and the matching results, we present in column (a) all the possible point-to-point correspondences determined with the estimated homography. There are 78 corresponding pairs among the 287 and 325 points detected in the first and the second image, respectively. After the first step of our matching algorithm that is the comparison with the similarity measure, we obtain 53 matches (29 correct and 24 incorrect). Next, we apply the additional verification based on the cross-correlation of affine normalized image patches. This verification rejects 10 matches (2 correct and 8 incorrect). The remaining 43 matches (27 correct and 16 incorrect) are displayed in column (b). Finally, there are 27 inliers to the robustly estimated homography, which are presented in column (c). Note, that there is a large perspective transformation between these images. The limited benefit of using the cross-correlation can be explained by a high similarity between different corners, which become similar after the affine rectification.

A second example is presented in figure 6.4. The images show a 3D scene with significant depth, and are taken from different viewpoints. This pair of images presents a larger change in viewpoint than the images in figure 6.7. There are 14 inliers to a robustly estimated fundamental matrix. In figure 6.5, we show the pair of images, for which our matching procedure fails. The failure is, however, not due to our detector, as the manually selected corresponding points show. It is, in fact, due to the differential descriptors, which are not sufficiently distinctive. Note that the corners of sharp or wide angles, of light or dark intensity are almost the same once normalized to be affine invariant. If there is no distinctive texture in the region around the corners there are too many mismatches and additional constraints, as for example semi-local constraints [30, 99, 119] should be applied. This also shows that a more robust and distinctive description is required to handle such significant viewpoint change.

6.3 Experimental results for image retrieval

In this section we present the experimental results of our approach in the context of image recognition. The database contains 5000 and 2000 images, for the scale and the affine invariant approach, respectively. We limited the number of images for the affine invariant detector in order to build the database in a reasonable time period. The images in the database are extracted from 16 hours of video sequence, which includes movies, sport events and news reports. Similar images are excluded by taking one image per 300 frames. Furthermore, the database contains one image from each of the test sequences. In order to reliably evaluate the performance of the recognition method we prepared a set of real image sequences. The images represent different outdoor and indoor scenes containing a

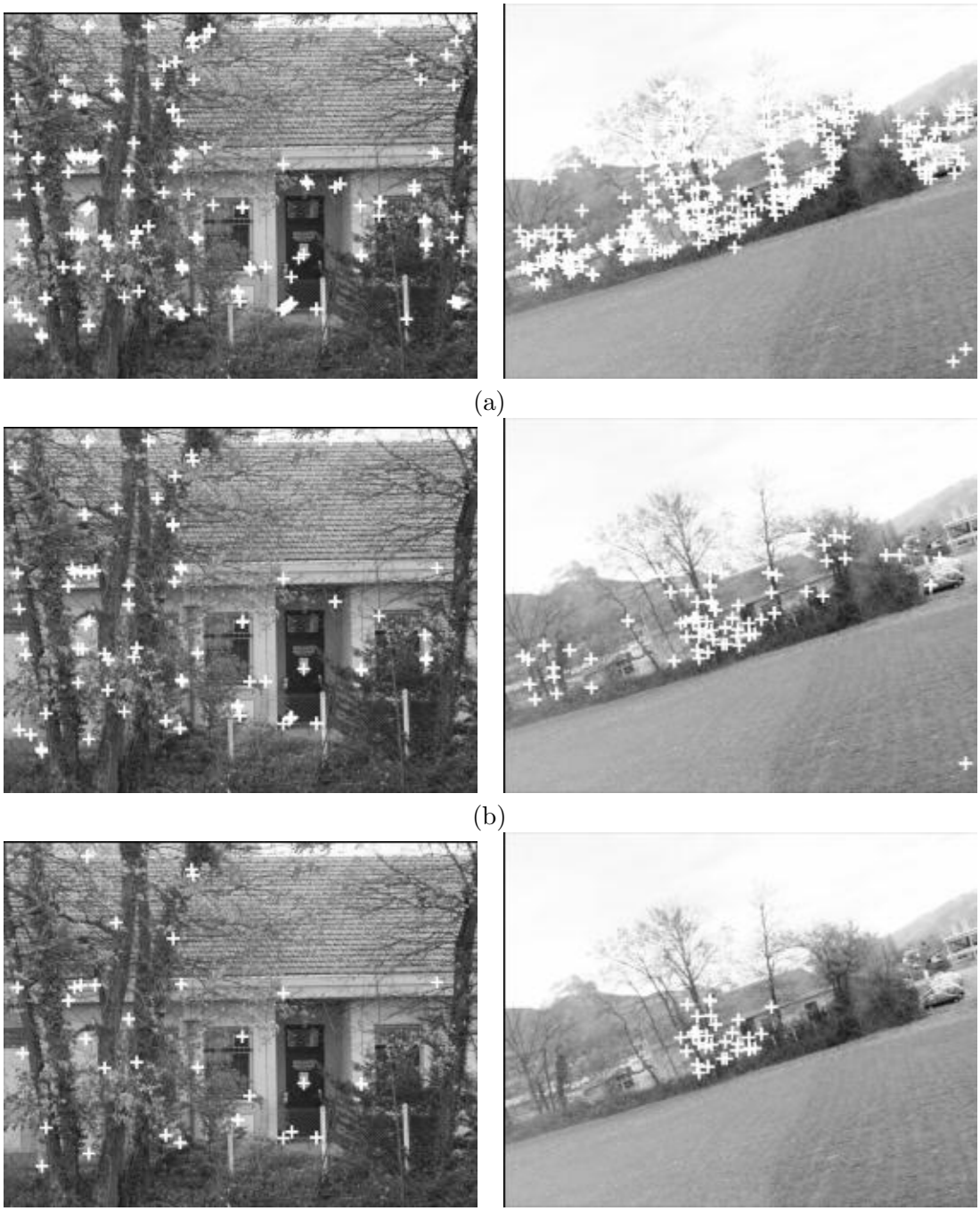


FIG. 6.1: *Robust matching: there are 190 and 213 points detected in the left and right images, respectively (a). 58 points are initially matched (b). There are 32 inliers to the estimated homography (c), all of which are correct. The estimated scale factor is* 4.9 *and the estimated rotation angle is* 19 *degrees.*

FIG. 6.2: *Example of images taken from different viewpoints. There are 14 inliers to a robustly estimated fundamental matrix, all of them are correct. The estimated scale factor is* 2.7.

wide variety of forms and textures. We estimated the real transformation between each pair of similar images, independently of our matching algorithm. This enabled us to reliably verify the results obtained with different approaches. The transformation is either the fundamental matrix or the homography. The image sequences are presented in annex A.3. There are 6 sequences presenting large scale changes and 6 sequences with significant viewpoint changes. The total number of descriptors in our database of 5000 images is 2 539 342.

In the following sections we present the retrieval results. We applied the approach described in section 6.1. Two algorithms are tested, one with the Harris-Laplace detector and the second one with Harris-Affine detector. The results for the two approaches are presented in sections 6.3.1 and 6.3.2, respectively.

6.3.1 Scale change problem

In this paragraph we show the results for retrieval from the database in the presence of scale changes up to a factor of 4.4. The database contains one image from each of our 6 test sequences presented in annex A.3. Each sequence contains the images from coarse to fine resolution. We introduced several fine resolution images and several of the coarse resolution to the database. The scale change is larger than 4.4 for some of the image pairs. The second row of figure 6.6 shows five images of the test sequences, which are in the database. The top row displays query images, for which the corresponding image in the database (second row) was correctly retrieved, that is it was the most similar one. The approximate scale factor is given in the third row. The changes between the image pairs (first and second row) include large changes in the focal length, for example 5.8 for the image pair (a). They also include significant changes in viewpoint, for example for pair (b). Furthermore, there is also considerable illumination change for image pair (e).

The retrieval results are presented in table 6.1. For each of the 6 test sequences, we

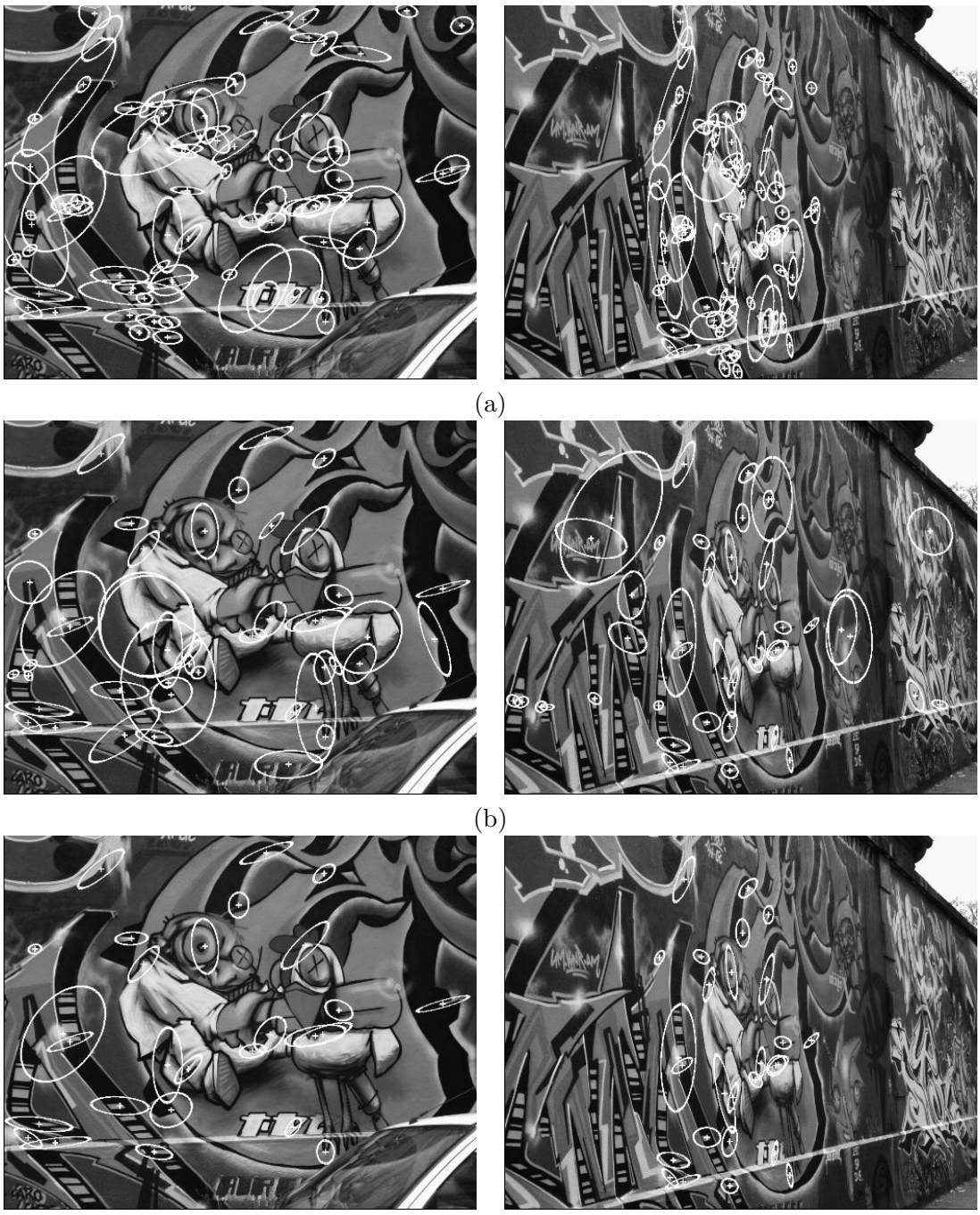


FIG. 6.3: *Robust matching: (a) There are 78 pairs of possible matches among the 287 and 325 detected points. (b) There are 43 point matches based on the descriptors and the cross-correlation score. 27 of these matches are correct. (c) There are 27 inliers to the estimated homography. All of them correct.*

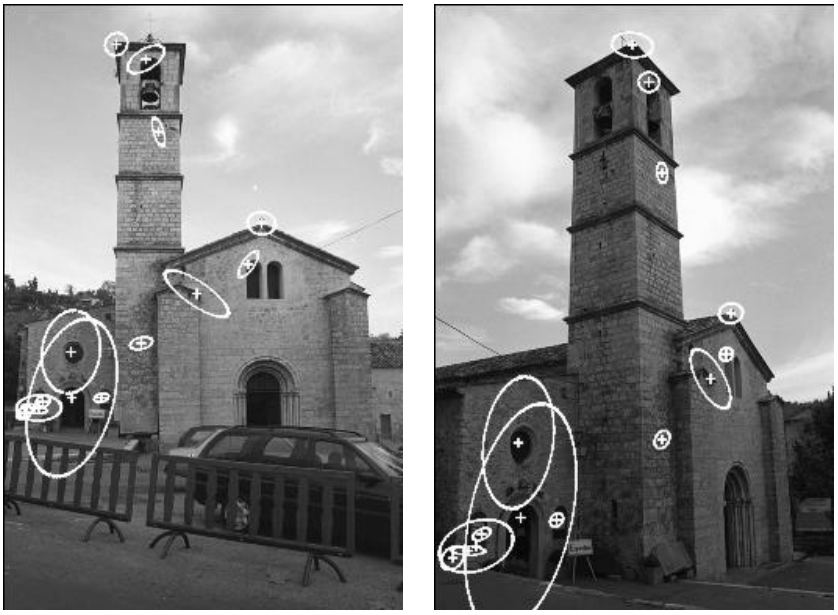


FIG. 6.4: *(a) Example of a 3D scene observed from significantly different viewpoint. There are 14 inliers to a robustly estimated fundamental matrix, all of them correct.*

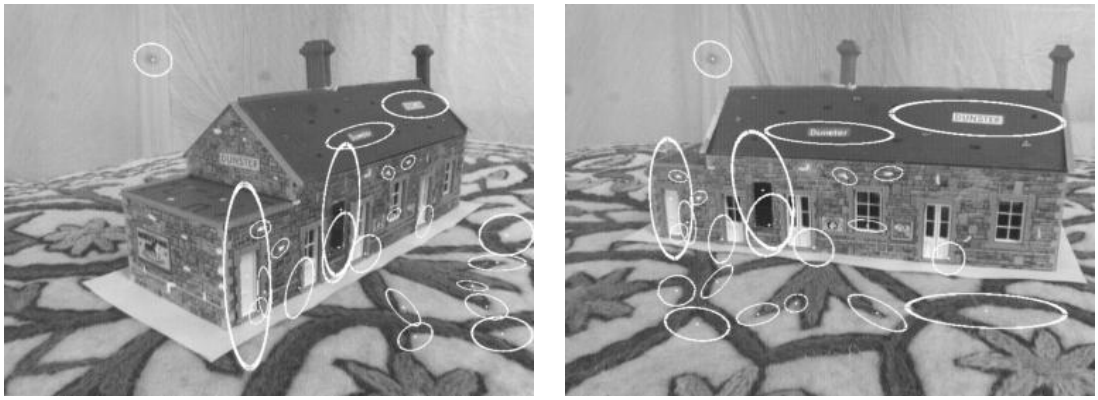


FIG. 6.5: *Example of an image pair, for which our method failed. There are, however, corresponding points, which we have selected manually.*

(a) 1/5.8 (b) 3.7 (c) 1/4.4 (d) 1/4.1 (e) 5.7

FIG. 6.6: *Correctly retrieved images. The top row shows some of the query images. The middle row shows the most similar images in the database, all of them are correct. The approximative scale factor between query image and database image is given in the bottom row.*

evaluated the performance at different scale factors (1.4 up to 4.4). For each scale factor, we obtained the number of the corresponding images retrieved as the most similar or among the five or ten most similar images. We can see that up to a scale factor of 4.4, the performance is very good. For the scale of 4.4, 2 images are correctly retrieved and 4 are among the 5 best matches. In the presence of uniform scale changes the Harris-Laplace detector performs better than the Harris-Affine. The Harris-Affine approach estimates the affine deformation of features, which rejects many points with correct scale and location but with highly anisotropic shape. The affine invariant points are also less distinctive (cf. section 5.3)

These results were obtained with 12 dimensional descriptors described in the paragraph 6.1.3. If we use derivatives up to order 3, that is 7 dimensional descriptors, the results degrade significantly. That confirms the usefulness of the fourth order derivatives.

scale factor	1.4		1.8		2.4		2.8		3.4		4.4	
detector	HL	HA	HL	HA	HL	HA	HL	HA	HL	HA	HL	HA
1	6	3	5	3	4	2	3	2	3	0	2	0
5	6	4	5	4	4	4	5	3	4	2	4	1
10	6	5	6	5	6	4	6	4	5	3	4	2

TAB. 6.1: *Indexing results for significant scale changes. The first row indicates the scale factor between test images. The second row indicates the interest point detector, the Harris-Laplace (HL) and the Harris-Affine (HA). The third row gives the number of correct retrievals, that is the corresponding image is retrieved as the most similar one. The fourth/fifth rows row gives the number of retrieved images which are among the 5/10 most similar images.*

6.3.2 Perspective deformation

The experiments for images with perspective changes were carried out on the database of 2000 images. Furthermore, we introduced to the database one image from each of our 6 test sequences presented in annex A.3. The second columns of figure 6.7 and figure 6.8 shows four of these images. The first column displays the query images, for which the corresponding image in the database (second column) was retrieved using the voting algorithm. Note the significant transformations including important scale changes between the query images and the images in the database. The images in figure 6.7 present a scene with considerable depth. There is also a large change in camera position and in lighting conditions. This image was retrieved as the second most similar one. There is a scale change of a factor of 3 between images 6.8(a). This image was retrieved as the third most similar one. Image pair 6.8(b) was taken with large changes in viewing angle. Image pair 6.8(c) combines a zoom change and a wide change in viewing angle. The images 6.8(b-c) obtained the highest similarity score in the database. The matched points displayed in the images are the inliers to a robustly estimated fundamental matrix or a homography between the query image and the correctly retrieved image in the database.

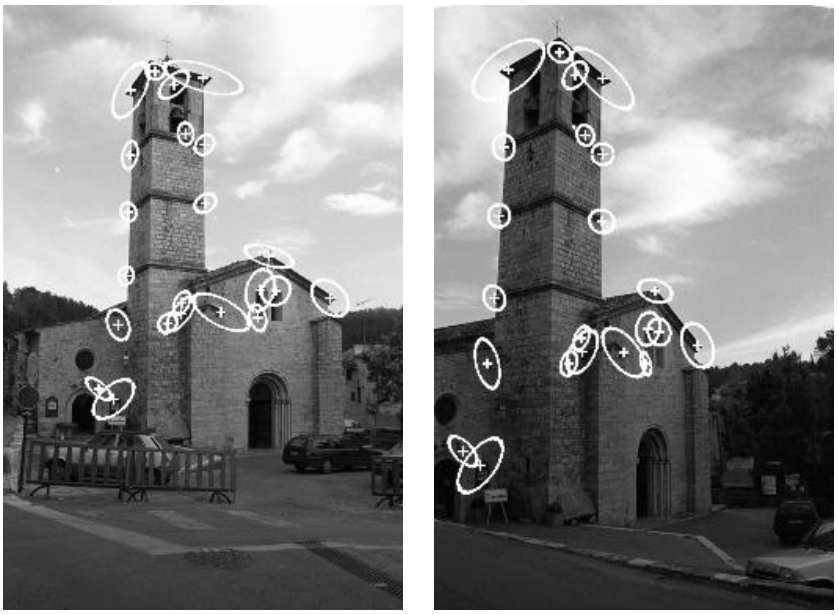
viewpoint angle	20°		30°		40°		50°		60°		70°	
detector	HL	HA	HL	HA	HL	HA	HL	HA	HL	HA	HL	HA
1	4	3	2	3	0	2	0	2	0	1	0	1
5	6	6	5	4	1	3	1	3	0	2	0	1
10	6	6	5	5	3	4	2	4	2	4	0	4

TAB. 6.2: *Indexing results for wide viewpoint changes. The first row indicates the viewpoint angle between test images. The second row indicates the interest point detector, the Harris-Laplace (HL) and the Harris-Affine (HA). The third row gives the number of correct retrievals, that is the corresponding image is retrieved as the most similar one. The fourth/fifth rows row gives the number of retrieved images which are among the 5/10 most similar images.*

The table 6.2 shows the results for images with perspective deformations indexed with the scale (HL) and the affine covariant features (HA). We can see that up to a viewpoint angle of 30 degrees, the performance of the methods is comparable. The performance of Harris-Laplace degrades significantly for wider angles. The detection algorithm is not adapted to affine changes of localization and point neighborhood. The Harris-Affine approach still provides a correct recognition for images with change in viewpoint angle of 70 degrees.

6.4 Discussion

In this chapter we have presented the results for matching and image recognition using invariant local features. Our algorithm is robust against arbitrary viewing conditions, occlusions and background clutter. The invariant feature detection considerably improves the



(a) (b)

FIG. 6.7: *Example of a 3D scene. The displayed matches are the inliers to a robustly estimated fundamental matrix between the query image (on the left) and the most similar image in the database (on the right). There are 22 matches. All of them are correct.*

classical approach to matching and retrieval. The main contribution to this approach is the interest point detection, which is invariant to significant affine transformations including large scale changes.

In this chapter we have shown that our approach enables excellent results to be obtained in both matching and indexing. The scale invariant points enable images with scale changes up to a factor of 4.4 to be reliably retrieved. However, these images present only weak perspective deformations. The scale change limit is determined by the range of scales applied to find the interest points. For images with a scale change of factor 4.4 there are 8 levels of 17 in the scale-space representation, which can be matched. The levels are located at finer scales for coarse resolution image and at large scales for fine resolution image. The complexity of the task is increased by the fact that there are fewer features at coarse scale levels. The points extracted from the 9 remaining levels, are useless. Thus, there is more than half the points that cannot be matched. The number of scale levels is limited by the size of the image. On the other hand, the neighborhood of points extracted at the finest scales is very small, therefore easily influenced by the noise.

The algorithm based on affine invariant points permits the correspondences between images with affine transformations to be found. The characteristic points are reliably extracted even from images representing significant perspective deformations including scale changes. The approach based on Harris-Affine detector gives stable results for wide changes in viewing angles but the method is less reliable for images with large scale changes, as we can see in the previous section. The affine invariance reduces the distinctiveness of points, thus the descriptors computed at the affine normalized points are less robust and

(a)

(b)

(c)

FIG. 6.8: *Correctly retrieved images. The displayed matches are the inliers to a robustly estimated fundamental matrix or homography between the query image (on the left) and the most similar image in the database (on the right). There are (a) 22 matches, (b) 34 matches, (c) 22 matches and (d) 33 matches. All of them are correct.*

are more often mismatched. All the conclusions concerning the scale invariant approach are also valid for the affine invariant approach. In the second case the complexity is increased because the scale changes are not uniform. The iterative method additionally rejects some points, which are detected for highly anisotropic structures. However, these points are useful in the case of uniform scaling. Thus, the scale invariant approach enables us to match and retrieve images with larger scale changes. We propose the Harris-Affine detector as a general solution to the affine problem and the Harris-Laplace detector for image transformations limited to scale changes with weak affine deformations. We can also use them both for unknown transformations.

To exceed the scale limit we have to make the approach more robust, as we cannot increase the resolution of images. An interpolation of images provides a partial solution to this problem because it does not increase the quantity of information. Therefore, we have to improve the description of local features and the process of finding similar points. Several possibilities are described in section 8.2.1.

Chapter 7

Recognition of an object class

THIS chapter presents a new method for detecting faces in a video sequence. When detecting complex objects the specific cues of their appearance must be taken into account. Therefore, the best results are obtained with methods developed for detecting one category of objects. The human face is an excellent example of such an object and the detection of faces is very important for analyzing video sequences. We propose a new approach to this problem, where detection is not limited to frontal views. The proposed method integrates the information of face content provided by the detector and the temporal information provided by the tracker. It simultaneously enables the detection and the tracking in a video sequence and provides better results than each of these approaches, when used separately.

In section 7.1 we introduce the problems related to face detection in a video sequence and present the existing solutions. In our algorithm, face detection is fully automatic and without loss of efficiency, it is a simplified version of the method developed by Schneiderman [110]. The wavelet based detector is presented in section 7.2. The temporal extension of this detector is based on the Condensation filter proposed by Isard [52] and can handle multiple faces, appearing/disappearing faces as well as changing pose and scale. This approach is explained in section 7.3. Experiments carried out on a large number of controlled movie sequences show a clear improvement in the results of frame-based detection (when the detector is applied to each frame of the video sequence). The comparative results are presented in section 7.4.

7.1 Introduction

In the following we explain our motivations for developing a method for face detection in a video sequence and we outline our approach to this problem. We then briefly describe the related approaches presented in literature.

7.1.1 Motivations

Face detection and tracking find applications in areas like video structuring, indexing and visual surveillance. These applications form an active area of research. If the objective is to identify an actor in a video clip [112] or to find the particular shot in the video sequence in which the actor is playing, then faces are the most important "basic units". Therefore a reliable detection and tracking of a face through a sequence is very important.

An approach to handling these issues could be *frame-based detection*, that is, to detect faces in each frame without taking into account the temporal information. Face detection requires determining the presence of a face and locating it within the image, by distinguishing it from all other objects and patterns present in the scene. This involves choosing an appropriate model for the face and segmenting the faces in an image. Models depend on the features and cues, which can be used for defining a face. An approach for face detection has to take into account the fact that there are different sources of variation of facial appearance like viewing geometry (pose), illumination (color, shadowing and self-shadowing), the imaging process (resolution, focus, imaging noise, perspective effects) and other factors like occlusion, shadowing and indirect illumination. The frame-based approach is useful when individual frames need to be processed as it completely neglects the fact that the visual information is contiguous in the sequence.

Another solution is *to use detection and tracking*, in which the face is detected in the first frame and followed through the sequence using tracking. In such an approach, tracking and detection are independent and information from only one source is used at a time. Thus, the available information is not entirely explored. This motivated us to develop a temporal approach for detection and to use it to locate and follow human faces in a video sequence.

We present a novel approach, which integrates detection and tracking into a unified framework. It uses the temporal relationships between frames to detect human faces in a video sequence, instead of detecting them in each frame independently. The proposed algorithm first detects regions of interest, which potentially contain faces by using detection probabilities. These probabilities are propagated over time using the Condensation filter and factored sampling for accumulation of probabilities, prediction and updating. The prediction of detection parameters, which are position, scale and "pose" guarantees the accuracy of accumulation as well as a continuous detection. The accumulation of detection probabilities over time, enables a face to be detected in subsequent frames even if it was not detected in the first frame. This leads to the method independent of thresholds, which are necessary in the case of a frame-by-frame detection procedure. We use two detectors, one for frontal faces and the other for profiles. These are combined to obtain the intermediate head pose. We can also handle the appearance and disappearance of faces by updating with the probabilities produced by the detection routine. We have developed a framework for prediction and update which propagates the probabilities of detection and the detection parameters over time. The need for updating-prediction requires the use of a filter. We have used the Condensation (CONditional dENsity propagaTION) algorithm proposed by Isard and Blake [52]. The probability distribution over all the detection parameters is represented by random samples. The distribution then evolves over time as the input data

/ observations change. When analyzing a scene with multiple faces, we are confronted with a scenario in which each face corresponds to one maximum. The Condensation filter does not make any assumption about the form of state density. It can therefore, represent non-Gaussian densities and can handle multiple modes. In addition, the present state of detection parameters are conditional to their past state, which is estimated from the data sequence by the filter. Condensation filter applies dynamic models and visual observations to propagate the random set over time, using the associated weights. Also *factored sampling* propagates samples with higher weights over time. This is required in our scenario as we need the faces with a higher probability score to be propagated over time. Also we need to be able to update faces on the basis of their appearance/disappearance from the scene. Thus, the Condensation filter together with the factored sampling is appropriate for our purpose. The Kalman filter [41] could also be used to handle the prediction and update procedure.

7.1.2 Related work

Most of the existing approaches are either based on detection [97, 102, 117, 110] or tracking [18, 46]. In the following we present some of these approaches.

Detection. The approaches for face detection can be categorized in different ways depending on the characteristic face structures or on the cues like color, shape etc., which are usually used for this purpose. On the basis of local structures, they can use a combination of features like eyes, nose and mouth and then constrain the problem using the spatial relations between them [15, 51]. These methods require multiple detectors for each of the features and a large number of parameters and spatial constraints need to be adjusted. Alternatively, face detection could be carried out by using the whole face [102, 117]. Although, this does not require the decomposition of the face into spatial features, partial occlusions are difficult to deal with. Variations in the face are handled by preprocessing and learning [126]. Also face detection methodologies can be categorized with respect to the image information used for detection - color [122], geometric shape [25] or motion information [46, 135]. These approaches suffer from the drawbacks of the specific cues, for example, skin color is sensitive to changes in lighting conditions and motion information may be affected by alternate motion in the video. In recent years, model based approaches to the interpretation of face images have been described. Face models based on appearance [57, 126], statistical models [87, 88, 110] and active shape models [62] have been successfully used in detection. These approaches can handle the type of faces on which they have been trained. However, most of the literature deals with frontal faces. Pose variation has been discussed in [43, 110] and they require the detector to be trained on profile views. Face detection approaches also vary with respect to the type of face model - 3D models obtained by reconstruction, which is always a sensitive process. It uses two views to obtain a model by setting up correspondences [114]. The models can also be constructed by learning process. Learning approaches can be based on neural networks [102], example based learning [117], PCA based [57, 126] and based on Support Vector Machines [94].

Tracking. The detection can be sufficient, when faces are located in a static scene. In a dynamic scene, as in a video sequence, the location and appearance of a face changes continuously, and tracking is required to follow a face through the sequence. In order to deal with the face changes over time in terms of changes in scale, position and to localize the search for the face, it is essential to exploit the temporal correlation between the parameters in consecutive frames. Tracking [12, 46, 101] exploits the temporal content of image sequences. There is a variety of trackers proposed for faces [11, 111] or general objects [18, 46, 133]. Face tracking can be divided into two categories 1) head tracking [10] and 2) facial feature tracking [32]. Tracking methods involve feature tracking (contours, points) [46, 47] or use information within the contours (color [101, 133], regions [46]). Birchfield [10] combines these approaches to obtain a tracker, which uses the elliptical contour fitted to the face and the color information inside. This can handle out-of-plane rotations and occlusions, but is unable to handle multiple faces and requires manual initialization. Face tracking, like detection, can also be categorized on the basis of the cues used : shape [25], color [111] and statistical models [31]. Approaches, which combine these cues, have also been developed. McKenna et.al [101] combine motion detection with an appearance based face model. Multiple person tracking was performed using multiple Kalman filters. Tracking involves prediction and update, for which filters like Kalman filter [41] and Condensation filter [52] have been used. Exemplar based tracking approaches have also been proposed [124], although only specific features of the face, e.g. lips have been tracked. Most of these tracking approaches suffer from the problem of initialization, as most trackers are initialized manually. In addition, they are usually unable to handle new faces appearing in the scene.

In recent years, approaches have been described to automate the process of face tracking by initializing the trackers with a face, which has been detected in the first frame [72]. In [72], faces are detected in the first frame by template matching and then tracked through the sequence which has been subdivided into shots. Alternatively, face tracking and detection can be combined by detecting facial features like lips, mouth, nostrils and eyes and by tracking them through the sequence, but this imposes the constraint that these features would need to be visible and therefore only frontal views of the face can be handled. Feris et. al [32] use a statistical skin color model to segment the face candidate regions in the image. The presence of a face is verified using eye detection. Then pupils, lip corners and nostrils are detected and tracked through the sequence. XVision [46] also handles face tracking by following eyes and mouth, although these need to be initialized. The framework of [76] can be used to track multiple faces, but it does not permit the addition of new faces.

In [18], an alternative method of tracking non-rigid objects has been proposed using Mean Shift which does not use the temporal information. The mean shift tracking is robust to partial occlusions, clutter, rotation in depth and changes in scale and camera position, although it cannot handle multiple detections. The mean shift iterations are used initially to find the target candidate that is most similar to a given target model. This detection phase is followed by tracking based on color changes. Detection is used only for initialization of the tracker and not for updating.

7.2 Face detector

This section describes our implementation of the face detector proposed by Schneiderman and Kanade [110]. The main difference relies in the local features used for describing faces. Without loss of detection performance we have limited the number of descriptors by selecting the most distinctive for the face. We also attribute significant weights for more prominent features, which improves the detection results. In section 7.2.1 we introduce the theory of the wavelet transform in the context of compact and exact image representation, which is necessary for description of objects; otherwise the dimensionality of the description is prohibitive in practice. Next, we explain in detail, our implementation of the wavelet based face detector. The first step of the algorithm is the representation of face attributes, which is detailed in section 7.2.2. Section 7.2.3 shows how we compute the probability of face appearance for frontal and profile detectors and how we combine these two probabilities to predict a face pose (cf.section 7.2.4). Finally, we outline the detection algorithm in section 7.2.5.

7.2.1 Introduction to wavelets

Many successful approaches based on a wavelet representation have been developed in the context of multi-resolution analysis and compression. Motivated by this fact and the excellent face detection results obtained with such an approach [110], we have implemented the wavelet based method. In the following we briefly describe the advantages of using wavelets for extracting and describing image features.

The general idea behind wavelets in image processing is simply to look at the wavelet coefficients as an alternative representation of the image. Instead of performing operations on pixels we can work with wavelet coefficients. This gives us the opportunity to take advantage of their multi-resolution structure and spatial-frequency localization. The accurate and compact representation by wavelet coefficients is much appreciated in the domain of signal compression. In the context of object recognition a rich and simultaneously compact description is very important for reliable and fast classification.

To show how the wavelets rapidly propagated to the image processing domain, after their appearance, we need to recall the major contributions. The wavelet transform introduced by Morlet in 1983 as an "ondelette" for analyzing seismic responses [91]. In 1984 Grossmann and Morlet [45] developed mathematical models of wavelets, which led to the "orthogonal wavelet basis" proposed by Meyer [83]. A fast algorithm for this transform adapted to image processing was proposed by Mallat [78] in 1989. Daubechies [24] developed orthonormal basis of compactly supported wavelets,and showed the relation between wavelets and FIR filters. The development of the bi-orthogonal wavelet theory, enabled the definition of different filter banks for analysis and synthesis.

The wavelet transform in two dimensions is often realized in a separable way, although the non-separable design of the transform for image processing provides a finer decomposition in scale and a better isotropy. In our algorithm we apply the separable implementation. Two filters are required for the decomposition. The first one is the low pass filter provided by a scaling function, so-called mother wavelet. The second one, high

pass filter, uses a function, called wavelet (cf. figure 7.2). The filtering operation is done by the convolution with wavelet coefficients. The algorithm presented in figure 7.1, called non-standard decomposition, is more efficient for computing than other implementations. Each step of the transform computes a quarter of the coefficients obtained by the previous step. This enables the multi-scale representation of images to be obtained in real time. A relation between the image size and the number of necessary operations is illustrated in figure 7.3. The wavelet decomposition is faster than Fourier or Gabor transforms. It also provides better localization simultaneously in space and in frequency, due to compactly supported wavelets, or in other words finite basis functions.

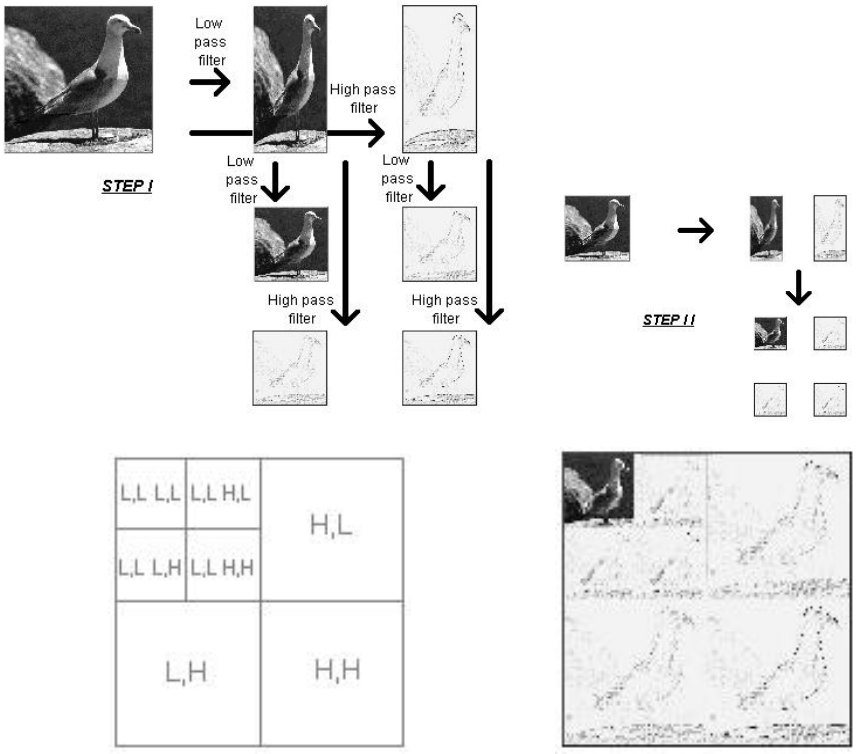


FIG. 7.1: *Wavelet transform with a pair of filters. Top row: Two steps of decomposition. Bottom row: Wavelet representation.*

To obtain a compact representation the wavelets must be adapted to the image signal. As a matter of fact, the four advantages of the wavelets for image analysis are the notion of multi-resolution, the relation with digital filters, the fast algorithm, and the linear phase. To avoid distortion in image processing the phase of digital filters must be linear, otherwise the edges in filtered images are distorted. It is also easier to implement the linear phase filters. A very important property is the regularity of the mother wavelet, which appears to be closely related to the regularity of the signal to be processed. Since images are generally smooth to the eye, with the exception of occasional edges, it is appropriate to use regular wavelets. The 7-3 biorthogonal wavelet (cf. figure 7.2) is a trade-off between the regularity,

visual effects on the image, and the complexity of the algorithm [6]. Another important criterion is the number of vanishing moments, in other words the oscillatory character of the wavelet. In practice, a wavelet with N vanishing moments enables the cancellation of all wavelet coefficients of a polynomial signal whose degree is less than N. Thus, the signal can be perfectly represented with very few wavelet coefficients. Unfortunately, not all these conditions can be satisfied, since there are no orthonormal linear phase short filters of a good regularity. Nonetheless, the highly important linear phase constraint corresponding to symmetrical wavelets is maintained by the biorthogonal spline 7-3 wavelet.

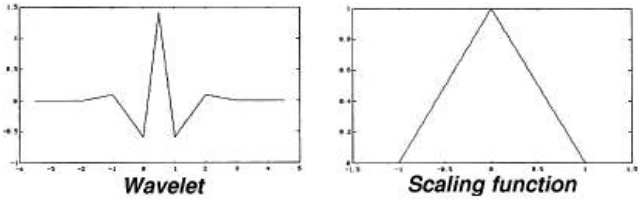


FIG. 7.2: *Biorthogonal 7-3 spline wavelets used in our algorithm.*

The excellent capacity of a compact representation can be illustrated by Comparing a histogram of pixel values with a histogram of coefficient values (cf. figure 7.3). One of the major drawbacks of wavelets in pattern recognition is that the transform is not translation invariant. In other words, when the input signal is shifted, the wavelet coefficients are not only shifted, but they also change their values. This effect can, however, be handled by an appropriate quantization of the coefficients [75]. Another problem is that the scale factor between successive resolution levels is 2. The information change is then very large in the context of recognition, therefore the intermediate scales have to be generated with the original image. There is an implementation, which permits a factor of $\sqrt{2}$ to be applied but it uses non separable and non-oriented filters [1].

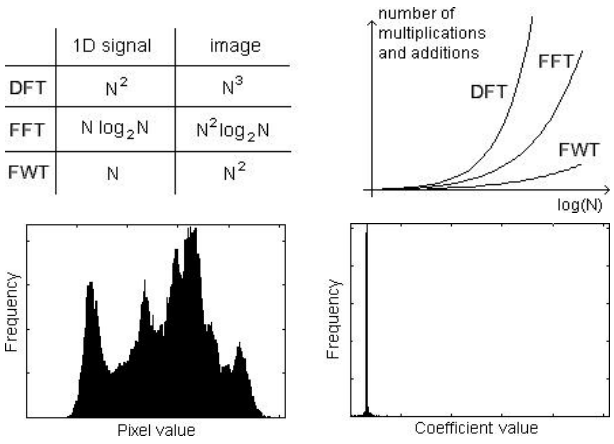
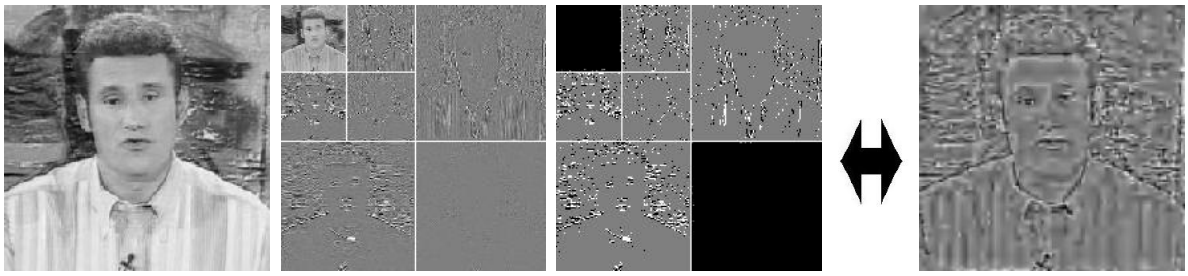


FIG. 7.3: *Advantages of the wavelet transform. Top row: Complexity of different transforms. Bottom row: Spectrum of a gray-level image and the wavelet representation.*

7.2.2 Appearance representation

Face decomposition. In the proposed approach an image decomposed with the wavelet transforms into several subbands representing different frequencies in horizontal, vertical and diagonal orientations at multiple resolutions. In figure 7.4(b) we show an example of a decomposed face. We use a representation obtained by a three step decomposition, although the figure presents only two levels for clarity. To obtain the invariance to light changes, skin color, and shadows we eliminate the low frequency band. We also eliminate the diagonal high frequencies as they convey less significant information (cf. figure 7.4(c)). Excellent work on wavelet quantization and coding was done by Vetterli and Kovacevic [75]. They show how to quantize and combine the coefficients to extract the essential visual information from wavelet representation. We follow this work and quantize each coefficient so as it takes one of the three possible values $[-1, 0, 1]$. Thus, keeping only the most important information about the form of local structures. Only the largest coefficients are assigned a non-zero value. However, we use different thresholds for each of the resolution levels. The threshold values were estimated empirically. Figure 7.4(d) shows a face reconstructed from the quantized coefficients.



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

FIG. 7.4: *(a) Input image. (b) Wavelet representation. (c) Quantized coefficients. (d) Reconstructed face.*

Visual attributes. Visual attributes should provide a representation of the face, which is jointly localized in space, frequency and orientation. After the quantization, each subset of coefficients is replaced by a discrete variable corresponding to the pattern represented by the coefficients. One pattern is represented by a combination of 8 coefficients, each represented by one of three quantization values, which makes $3^8 = 6\ 561$ possible codes for a given set of 8 coefficients. The 8 coefficients of one pattern are provided by different frequency and orientation bands. We use 12 different coefficient combinations. These codes are additionally combined with their location within an analyzed image frame. Each analyzed frame is divided into 16 square regions defined by the coordinates (x, y). This additionally increases the number of possible values to $16 \times 6\,561 = 104\,976$. Figure 7.5 shows the applied coefficient combinations. Given a location within the analyzed image window we compute the corresponding locations in the subbands of wavelet representation. The 8 coefficients are taken form different subbands LH and HL as showed in figure 7.5(b). The

visual attributes are therefore based on combinations of quantized wavelet coefficients at different positions and in different frequency bands. This provides a rich face representation and enables fine features like eyes, nose, mouth, ears, together with their spatial relations to be extracted and encoded.

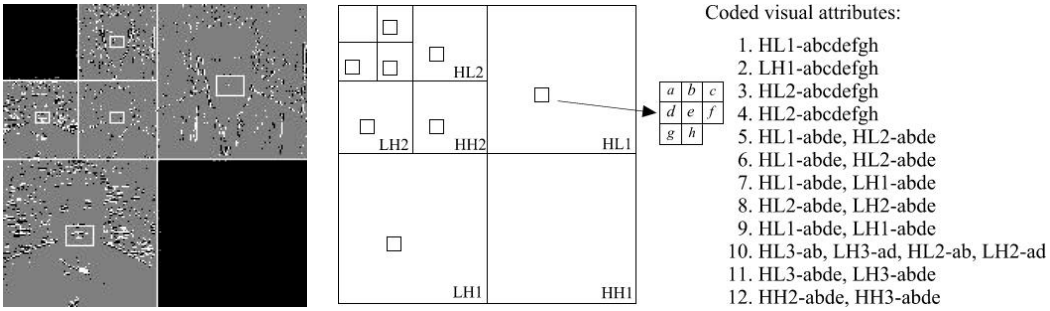


FIG. 7.5: *(a) Quantized coefficients. (b) Coded wavelet coefficients.*

Classification models. The classification models represent statistics of both face or non-face appearance. The statistics are based on the joint probabilities of visual attributes. The statistical distribution of each of the coded attributes is represented by a histogram of 104 976 bins. Each of the coefficient codes corresponds to one histogram bin. The probability of a bin is determined by the frequency of appearance of corresponding attributes in different images presenting faces (face model) and other scenes (non-face model).

We used about 150 different faces to build the frontal face model and 200 faces to build the profile model. The exact location of the eyes and the mouth, was determined manually. This enabled us to normalize faces with respect to the size and orientation. We use 5 different rotation angles uniformly distributed within the range of $(-15, +15)$ degrees with respect to the original face orientation. The size of smoothing Gaussian kernel was computed with $\sigma = 0.6 \cdot 1.2^n$, where $n = 0, \ldots 5$. To increase the training set, we created several smoothed and rotated versions for each face. If a code is present at a given position the corresponding bin value is increased. Finally, to obtain the normalized histogram each bin value is divided by a sum of all bins. This gives an approximated probability of appearance of a face attribute $P_a(code(x, y), (x, y)|face)$. Next, each bin with zero-probability was set to a value of 100 times less than the minimal bin value to remove the 0 probabilities.

It is more difficult to build a model for all image structures, which are different from the human face. Therefore, the training procedure is more complex. The non-face model was initially acquired from 200 images containing no faces. We then applied the detection procedure to the same non-face images. This gave rise to a set of false positives. In the second stage we acquired the non-face model from the false positive regions that gave the highest detection responses. The procedure was repeated for the frontal faces and the profile faces separately until we obtained satisfying recognition results. The training set for

face and non-face images was collected from the Internet and none of the test sequences were included in this training set. Thus, we obtained an approximated probability of appearance of non-face attributes $P_a(code(x,y),(x,y)|non-face)$. The profile detector was trained on images with faces turned at an angle of 45 degrees or more.

Color. We did not use skin color information for our detection algorithm, as our experimental results showed that the illumination can significantly change skin color. We verified several color representations to build a skin model as Lab, Luv, HSV and RGB. To build the model we used the dimensions ab, uv, SV and rg, where $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$. The distribution of the skin color was represented by a 2 dimensional histogram where each bin represented the frequency of appearance of a given color in the skin regions in our training set of faces. This model represents the information, common for different skins colors i.e. white, yellow, black. The most dispersed distribution was the one estimated for HSV color space. This space seems to be less adapted to represent the characteristic cues of the skin, although quantitative comparison would be valuable. An example of skin color detected with our rg model is presented in figure 7.6. The image was generated with thresholded bin probabilities of corresponding pixel colors. However, in professional videos skin color information is unreliable as artificial illumination completely changes the skin color.

FIG. 7.6: *Skin color detected with a model build in rg color space. The image displays thresholded probabilities of corresponding pixel colors.*

7.2.3 Probability score

The detector response is computed using a weighted combination of 12 visual attributes, evaluated on 16 regions of the windowed input image at a given scale:

$$R(Face|I,x,y,s) = \sum_{a=1}^{12} \sum_{(x,y)}^{16} \log \frac{w_a P_a(code(x,y),(x,y)|object)}{P_a(code(x,y),(x,y)|non-object)} \tag{7.1}$$

The logarithm of the probability ratio accumulates the responses, but significantly slows down the computation.

To obtain equivalent responses for each attribute, we use different weights w_a. The weights are estimated with the training data. This significantly minimizes the classification

error. The idea is to assign larger weights to the significant attributes. Given a set of 50 images with known location and size of each face, we varied one parameter at a time to obtain the best classification results. More sophisticated methods can be applied to find the optimal set of weights [100].

In order to detect frontal and profile views we use two detectors. The responses of the frontal and profile detectors are then $R_f(Face|I,x,y,s)$ and $R_p(Face|I,x,y,s)$, respectively. The coordinate x,y is the position at which the detector is applied and s is the scale. These responses are then normalized to lie between 0 and 1 to give the corresponding frontal and profile probabilities, $P_f(Face|I,x,y,s)$ and $P_p(Face|I,x,y,s)$. For simplicity, we denote the responses as R_f and R_p and the probabilities by P_f and P_p.

Figure 7.7(a) shows the probability map for our face example. Note that a displacement by one coefficient in the resolution level LH2 results in the displacement of 4 pixels in the original image. Thus, the probability in the displayed image does not change smoothly. To handle this problem we smooth the probability image with a Gaussian. We then obtain a probability map for an input image, where the local maxima correspond to potential face locations. If the value of a maximum is over a threshold, a face is assumed to be present.

(a) (b) (c)

FIG. 7.7: *(a) Probability map (b) Smoothed probability (c) Output image*

7.2.4 Pose representation

Pose changes are the most important source of variation in face appearance. The detection of faces from any viewpoint requires training a multi-view face model. However, the training on all possible views makes the model much less distinctive, which entails many false classifications. In order to detect faces from any viewpoint, we have used two detectors. One was trained to detected frontal views, the other to detect profiles. Note that the profile detector is also applied to the mirror-reversed image. In each location, we only keep the higher response of both images to indicate the probability of profile. We then apply a method to estimate the actual pose using only the frontal and the profile probability.

In figure 7.8 we can see that the probability of the frontal view detector decreases, as the face turns sideways and vice versa. Since the probabilities for the frontal and profile views have been obtained using different classification models, their responses are not comparable. Note the the sum of the probabilities is not equal 1. The difference is, in

fact, equal to the probability that the image does not contain a face. Also the maximal response for the two detectors, when applied to an intermediate face pose, has slightly different localization. We have estimated the displacement vector from our training set. Difference in localization of the maxima is illustrated in the first row of figure 7.8. On the basis of the experimental validation that the frontal responses decrease with the face turning away and profile responses increase, we normalize the responses in order to make them comparable. That is, the responses are normalized to a range of values between 0 and 255. We find that a greater number of pixels give a response between 200 and 255 for profile detector. The responses are then modified so that the responses between 200 and 255 get spaced out to between 150 - 255. In this way they become comparable to frontal detector responses. These can then be divided by 255 to obtain the frontal and profile probabilities. Therefore we are able to combine the probabilities P_f and P_p to find the approximate head pose. In the context of a sequence of images, the observed pose at time t is given by:

$$\lambda_t = \frac{P_p(face|I_t, x_t, y_t, s_t)}{P_f(face|I_t, x_t, y_t, s_t) + P_p(face|I_t, x_t, y_t, s_t)}$$

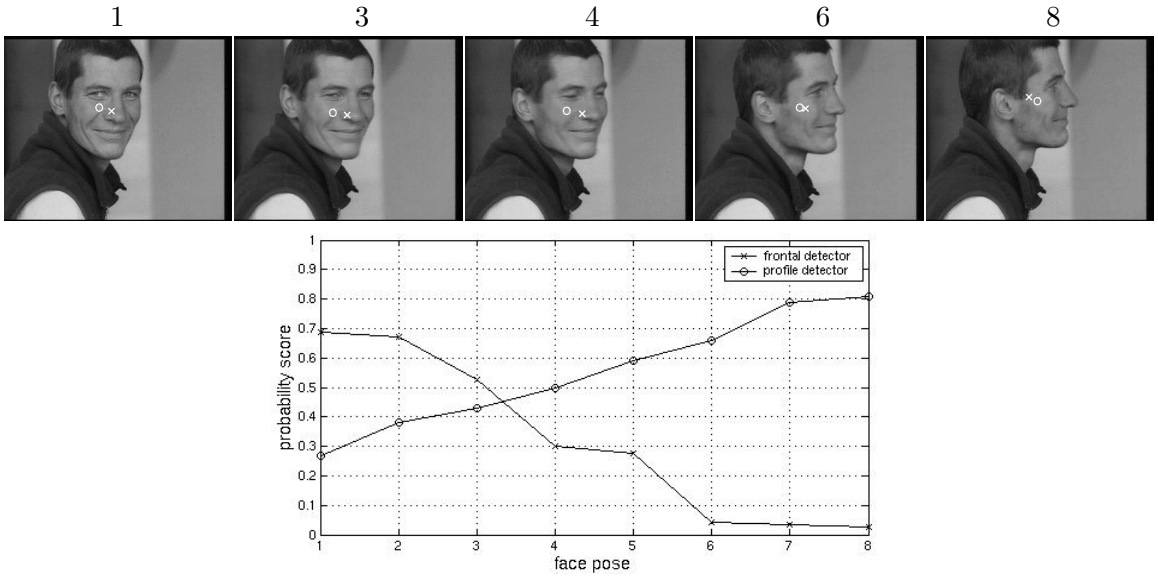


FIG. 7.8: *Varying face pose. This first row shows frames 1, 3, 4, 6 and 8. The local maxima for the frontal and profile detectors are displayed: a cross indicates the location for the frontal detector, a circle the location for the profile detector. The third row shows the probability scores at the respective maxima for frontal and profile detection as a function of the frame number. The probability of the frontal detector decreases as the face turns and the probability of the profile detector increases.*

7.2.5 Detection algorithm

The detector finds faces at a given location and a given scale. It has to be applied to all image positions and at different scales. The ratio between two consecutive scale levels is an empirically chosen constant and equals 1.2. The model represents a face fitted into a bounding box of size 64×64 pixels. Therefore, the region being classified must be of the same size. As a matter of fact, the evaluated region is smaller in different resolution levels of the wavelet representation. We interpolate, smooth and sample the input image to obtain a scaled version. Next, a wavelet representation is computed and the coefficients are quantized for the entire image. To classify a given image window we apply the routine, which selects the corresponding coefficients at different frequencies and orientations. It computes the codes also using the coefficient locations in the window coordinates. Next, a probability map is computed and local maxima are determined. A face is actually detected if the maximum is above a fixed threshold. This provides a list of potential faces in the image. Since a maximum can exist over a range of scales, we remove collisions by choosing and maintaining the one with the highest response. Given a maximum we look for all other maxima at every scale within a window of the size determined by the scale. If there is a maximum with a higher response the investigated face is removed from the list. Finally, we obtain a list of locations, scales, and appearance probabilities of faces. Several examples of face detection in single images are presented in section 7.4.3.

7.3 Face detection in a video sequence

In the following we propose a method for identifying multiple faces in a video sequence. We first present a general idea of using the temporal information represented by a sequence of images and we then explain the implementation details.

7.3.1 Temporal approach

Initially the detector described in the previous section is used to associate a probability of detection with each pixel, in every frame of the sequence. A probability value is also computed for different scales and for two different views. We use two detectors, one for frontal and the other for profile views. By "pose" we mean the frontal and the profile views and the positions of the face in between these two views. The parameters, which characterize the face are therefore the location, the scale and the pose. These form our detection parameters, which have to be propagated over time. All the parameters could be computed using a frame-by-frame detection, but the detector response can decrease for different reasons (occlusions, lighting conditions, face pose). Without any additional information these responses can easily be rejected even if they still indicate the presence of a face. This is due to a fixed threshold. If the threshold is too low there are many false detections. The evaluation for several preselected scales can be inaccurate in the sense that the actual size of the face is difficult to determine when it does not match with one of the applied scales. The evaluation for every candidate region of the coefficient representation is the most time consuming part of the algorithm. Also some intermediate poses are not

detected by the profile detector.

Thus, we exploit the temporal correlation of the detection parameters in a video sequence by adopting a prediction-update model to determine the parameters for each frame. The information about the predicted face location and scale can significantly accelerate the algorithm and continuously detect faces even in the frames where the threshold based detector fails. This also helps accumulate the probabilities of detection over a sequence to obtain a stable detection over time, which is independent of thresholds. We use the Condensation filter [52] and factored sampling to propagate the parameters over time. The tracker integrates the responses from the two detectors with the temporal information, to detect all intermediate poses.

The local maxima of the probability distribution produced by the detectors are used to initialize the procedure. The local maxima indicate the possible occurrence of faces in one of the scales. Samples are picked with a Gaussian centered at each maximum. These samples are then propagated across the sequence. The prediction is based on a zero order model for position, scale and pose. The update at each stage uses the probability map generated by the detection routine. Thus, our proposed procedure is divided into two phases. The first phase is the detection, which produces the probabilities for each image location, scale and viewpoint. This is described in the previous section. The second phase is the prediction and update stage, which predicts the detection parameters and uses the probabilities to track the face through the sequence. Temporal propagation is described in the next section.

The combination of the two detectors allows us to estimate an approximate pose. This approximation is sufficient for prediction/update. If we use only one of the detectors, we cannot predict in the case of decreasing probability whether the face is turning or disappearing. The initialization indicates whether the face was initially in the frontal position or profile. The tracker then follows this up by prediction. At each stage we check whether the face is turning in the predicted direction and the responses from the detectors correspond to our model (cf. section 7.2.4). This will be explained in greater detail in the next section.

7.3.2 Adapted condensation algorithm

In the following, we show how to incorporate the temporal information between the frames with the detection information so that we can track the face through the sequence. The detection parameters, which define an appearance of a face are:

- (x, y) : location. We examine every image location.
- s : scale at which the face occurs. We apply a discrete range of scales which have been empirically chosen.
- θ : face pose. The parameter θ can take any value between 0 and 90 and it indicates the corresponding face pose. 0 corresponds to the frontal face and 90 to the profile faces. We do not distinguish between the two profile views.

The *state* at time t, s_t is defined to be a vector of parameters

$$s_t = (x_t, y_t, s_t, \theta_t)$$

The *observations* at each stage are the probability values computed by the detector, which is explained in section 7.2. The probabilities $P(face|I_t, x_t, y_t, s_t)$ are the values associated with each location in the image. Each scale is represented by an image. There are two different probabilities associated with the two "poses" of the head; $P_f(face|I, x, y, s)$ corresponding to the frontal face detector and $P_p(face|I, x, y, s)$ corresponding to the profile detector. The observation z_t is then given by:

$$z_t = (P_f(face|I, x, y, s), \quad P_p(face|I, x, y, s))$$

These probability values determine the likelihood of observations and the conditional probability $P(z_t|s_t)$, that is the probability of observation z_t given the state s_t. We denote the probabilities as $P(x, y, s)$ with the suffixes for the frontal and side views used only if we need to distinguish them. Given this conditional probability distribution, a discrete representation of the entire probability distribution can be constructed over the possible states. Our algorithm is divided into 4 steps:

Step I: Initialization. We initialize the algorithm using the local maxima provided by the detector applied at the initial frame of the video sequence. Each of the maxima is propagated separately. Contrary to the single image detection the threshold is not used. Maxima corresponding to non-faces are eliminated over time.

1. We sample the probability around the maxima (x, y). We keep the scale fixed in the initial sampling to consolidate the maxima over the scale. Initially, if the samples are picked over very different scales, there is a chance of losing the maxima. The samples could be picked randomly one scale higher or lower than the scale at which the maxima occurs. We select 300 samples around each maximum.

2. We then choose the probabilities corresponding to the samples from the respective location of the front and profile faces. The pose is initialized with:

$$\theta_0 = \frac{P_p}{P_f + P_p} \times 90$$

The corresponding total probability P is then given by

$$\begin{aligned} P(face|I_0, x_0, y_0, s_0, \theta_0) &= P_f(face|I_0, x_0, y_0, s_0) \text{ if } 0 < \theta_0 \leq 45 \\ &= P_p(face|I_0, x_0, y_0, s_0) \text{ if } 45 < \theta_0 \leq 90 \end{aligned}$$

The set of probabilities are normalized to produce the weights

$$\Pi_i = \frac{P_i}{\Sigma_{i=1}^{S} P_i}$$

where S is the total number of samples, which are being propagated.

The sample states and the weights are used to predict the probability distribution at the next time instant. The next three stages set up a new probability distribution at time t given the distribution at time $t-1$.

Step II: Selection. We use factored sampling [52] to sample the states at stage $t-1$. These are then used for propagation to the next time instant. The sampling method selects the states with respect to the associated weights. Samples, which have the highest weights are propagated. Samples with high response are propagated, while those with lower responses are eliminated. Thus, we proceed to eliminate the non-faces and propagate the faces.

Step III: Prediction. We use a zero order temporal model for the prediction of a new state

$$\begin{aligned} x_t &= x_{t-1} + N(\sigma_x) \\ y_t &= y_{t-1} + N(\sigma_y) \\ s_t &= s_{t-1}(1.2)^k \text{ with } k \in N(\sigma_s) \\ \theta_t &= \theta_{t-1} + N(\sigma_\theta) \end{aligned}$$

In the above, the scaling factor of 1.2 has been empirically chosen and gives the best results. For our experiments we have set the parameters σ_x and σ_y to 5 pixel, σ_s to 0.5 and σ_θ to 1 degree.

The prediction approximates the conditional probability of the present state given the previous state.

Step IV: Updating. The probabilities of the predicted states are combined to obtain the probability associated with each state:

$$P(face|I_t, x_t, y_t, s_t, \theta_t) = max(f(\lambda_t, \theta_t)P_f(face|I_t, x_t, y_t, s_t), f(\lambda_t, \theta_t)P_p(face|I_t, x_t, y_t, s_t))$$

where λ_t is the observed pose angle at time t:

$$\lambda_t = \frac{P_p(face|I_t, x_t, y_t, s_t)}{P_f(face|I_t, x_t, y_t, s_t, \theta_t) + P_p(face|I_t, x_t, y_t, s_t, \theta_t)} \tag{7.2}$$

and θ_t is the predicted probability obtained from Step III. The function f is a function of the difference between the observed and the predicted pose angle. If the angles are the same, then the response should be high. The higher the difference between the angles, the lower the response should be, since we would like the incorrectly predicted sample to be eliminated in the consecutive time steps. Thus, f can be any linear function, which satisfies these conditions. The values of θ_t and λ_t can change in the range between 0 and 90. Therefore, in our case, we choose f to be

$$f(\lambda_t, \theta_t) = 1 - \frac{|\lambda_t - \theta_t|}{90}$$

As seen, the value of f is 1 when $\lambda_t = \theta_t$ and is 0 when it is 90, that is, the maximum difference that they can have.

then normalized to obtain the total probability $P(face|I_t, x_t, y_t, s_t, \theta_t)$ associated with the

state. We understand here that the responses from the two detectors are normalized in order to make them comparable.

If the likelihood of all points goes to zero, a new prediction is obtained from Step III. To consolidate the propagated samples at each stage we find the weighted mean and variance of the states:

$$Mean(S) = \Sigma_{i=1}^{N} \Pi_i S_i, \quad Var(S) = \Sigma_{i=1}^{N} \Pi_i S_i^2 - Mean(S)^2$$

where N is the number of samples. The mean values indicate the localization of the face. The average value of scale stabilizes it over time. The variance is stable over time for the faces, but increases, that is gets diffused over time for non-faces. Also in the case of non-faces, the corresponding probabilities go to zero.

Appearance and disappearance of faces. To handle the situation when new faces appear in the scene, we update the set of local maxima at every nth frame. This permits new faces which appear in the sequence to be detected. The faces, for which the variance increases and gets diffused over the image, are abandoned. Alternatively, regions in the image may show constant low probability for all the points, which are propagated. The detection parameters are maintained for each face for five consecutive frames. If the analyzed face continues to indicate low probability, which does not diffuse over time, it indicates a false maxima and is deleted from the list of faces.

7.4 Experimental results

We present extensive experimentation carried out on a large number of sequences, which includes sequences taken in a controlled environment as well as sequences taken from movies. The aim of the experiments is to compare the frame-by-frame detection method with the "temporal" approach. Experiments have been carried out on video sequences in which multiple faces appear or disappear, change location, scale and pose.

We consider that a detection is correct if the location of the bounding box indicates a presence of a face. False detections appear when a bounding box is drawn without the presence of a face. A face is incorrectly detected if there is no response from the detector or if the location or the size of the bounding box do not fit to the face.

7.4.1 Single frame detection

The detection algorithm presented in section 7.2, was applied to each frame of the video sequences. The scale at which the face is detected is used to determine the size of the bounding box. For the video sequences we have chosen the threshold, which gives the best classification results, that is it minimizes the ratio between false detections and missing detections.

Results for individual images are shown in figures 7.9, 7.10, Figure 7.9 shows faces with occlusions and uneven illumination. There are three partially occluded faces, for which the probability was below the threshold. We can see that the frontal detector is robust to small variations in pose (cf. figure 7.10(a)) and low quality images. Figures 7.10(d)-(f) present

the results for frontal detector. There are many false detections due to low classification threshold and the bounding boxes do not fit exactly to the profile faces. The correct probability maxima are not as distinctive as the ones for frontal detector. These might be due to less distinctive classification models, which accumulate the statistical appearance of local image patterns of profile views. Note that the profile patterns are less distinctive from the background than the frontal face patterns.

7.4.2 Temporal detection

The temporal framework explained in section 7.3 is applied to each of the video sequences. The most significant local maxima associated with the first frame of the sequence are used to initialize the procedure. Without loss of generality, we have used the 4 strongest maxima for our experiments. These maxima are then propagated through the image sequence. Weighted mean and variance is computed for each frame using all the samples. The mean value for location and scale are used to display the bounding box. The probabilities of false maxima and the samples picked around them decrease to zero over time due to change in illumination and camera position, and are therefore eliminated. The weighted mean and variance for such maxima decrease to zero.

7.4.3 Observations

In the following examples we compare the results provided by the "temporal" approach the results obtained with the frame-based detector. We applied the frame-based detection and the temporal approach to several sequences. Next, we evaluated every frame of the sequence that is 1020 images.

In the frame-by-frame detections, which are displayed in the first rows of figures 7.11 - 7.14, we obtained 11.2% false detections and 19.4% incorrect detections. The temporal detection results are displayed in the second rows of figures 7.11 - 7.14 and all are correct. The frame numbers are given above the images.

Frontal detectors. We observe that for the temporal approach there are no missing detections, no false detections and the detection results are continuous throughout the sequence. There are no false detections because the false maxima, which appear in the first frame, are subsequently eliminated. These maxima are picked for initialization and are propagated/increased over time, but their probability decreases to zero in subsequent frames. Therefore, we can eliminate them from all frames (cf. figure 7.11). The frame-by-frame detection fails, because the probability score is below the detection threshold. Furthermore, for the temporal approach, there are no incorrect detections as the face is continuously tracked through the sequence. The average value of the parameters over all the samples smoothes the results. Consequently, the position and the size of the bounding boxes (scale) vary smoothly over time. The trembling effect of bounding boxes is visible in the sequence obtained by frame-based detection (cf. figure 7.12). The temporal approach is able to detect the disappearance and appearance of faces before the detector, because it keeps track on the faces (cf. figure 7.13). The faces continue to be tracked even in the case of a head movement and out-of-plane rotation, while the frame-by-frame detection

loses some of these faces, as the probability score goes below a threshold (cf. figure 7.14). The multi-scale approach combined with the Condensation filter enables stable detection results to be obtained also for zoom sequences (cf. figure 7.15).

Combined frontal and profile detector. In figure 7.16, we show the results of applying the frontal and profile detector to a sequence in which the face is turning away from the camera. The face is not detected by the frontal detector when its left part disappears due to rotation of the head. On the other hand the frontal faces are incorrectly detected by the profile detector. This indicates that a single detector is insufficient. In the temporal approach, the responses from the two detectors are combined and the face is continuously tracked through the sequence. The face is correctly detected in all frames.

FIG. 7.9: *Detection results for the frontal face detector. Note that the faces are correctly detected despite shadows and occlusions.*

7.5 Discussion

In this part of the dissertation we have presented our implementation of a face detector based on the distribution of local appearance of a face. This approach represents a compromise between methods, which are efficient in computation time but not robust, and the methods, which give reliable results, but are very time consuming. Detection based on skin color can be done in real time but it is not robust to illumination changes, and gives false alarms for other parts of human body. Thus, we cannot base the recognition only on color information. On the other hand there are many methods, which require complex, time consuming computations like the approaches based on Gaussian mixtures.

In the presented approach, face attributes are captured by combinations of wavelet coefficients. The wavelet representation is very compact, therefore we use many different

(a) (b) (c)

(d) (e) (f)

FIG. 7.10: *Detection results: (a) - (c) - Frontal face detector , there are three missing detections. (d) - (f) - Profile face detector. Note some false responses of the profile detector.*

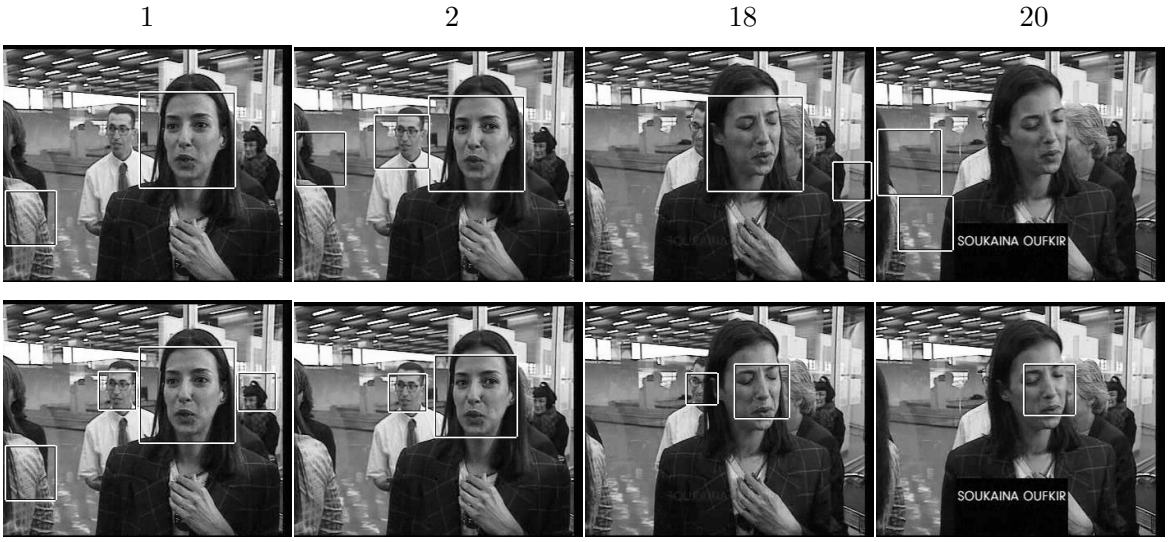


FIG. 7.11: *The face of the small man was not detected by the simple detector in frames 1, 18 and 20 (top row). The corresponding images in the second row show that the man is detected by the "temporal" detector. Note that the small man is not detected in the first frame. There is, however, a local maximum with a low probability at the location of the face. In the case of the "temporal" detector, this maximum is sampled for initialization and propagated/increased over time. The frame-by-frame detection fails, because the probability score is below the detection threshold. There are two non-faces detected for frame 1. The "temporal" detector uses these maxima for initialization, but their probability goes to zero in subsequent frames and are therefore eliminated.*

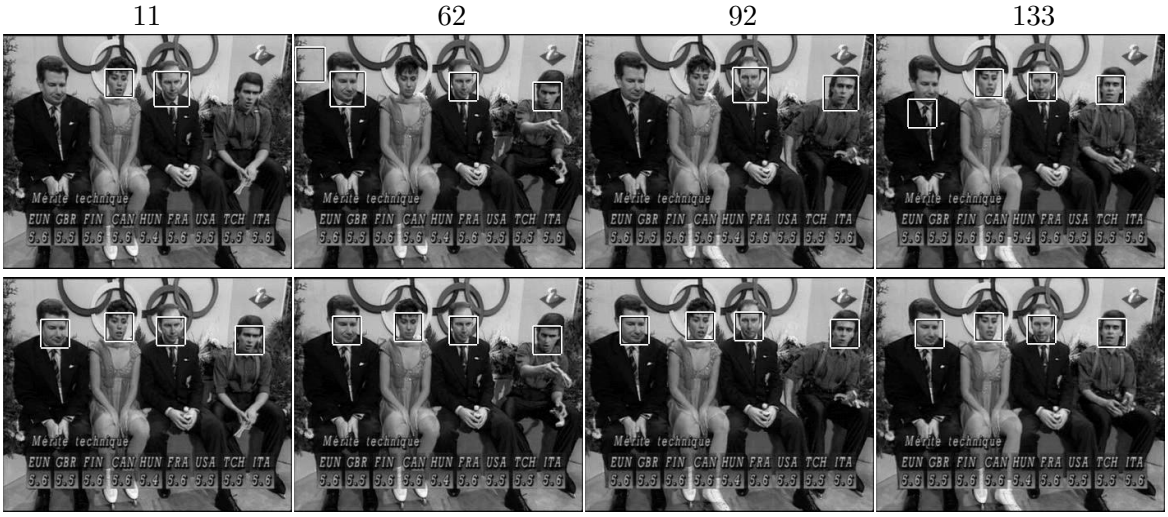


FIG. 7.12: *The first row shows faces detected with the frame-by-frame detector. The second row shows results of the "temporal" detector. The results of the "temporal" detector improve significantly over frame-by-frame detection.*

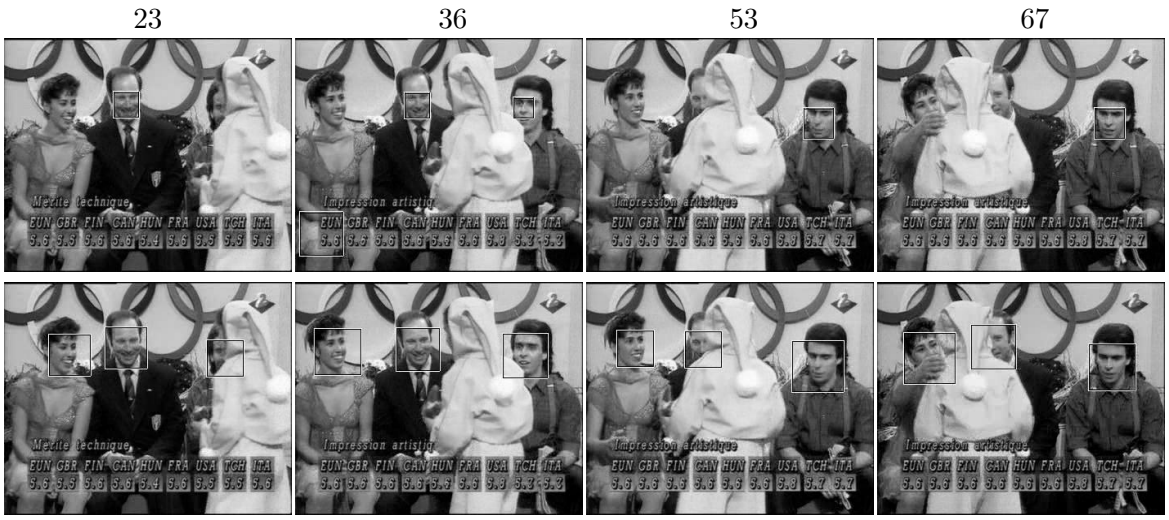


FIG. 7.13: *Detectors applied to a sequence, in which faces get occluded and reappear. The first row shows faces detected with the frame-by-frame detector. The second row shows results of the "temporal" detector. The faces are occluded by the pedestrian and then reappear in the sequence. Frames 53 and 67 show that the temporal approach is able to detect the disappearance and appearance before the detector. This is because the temporal approach keeps track of the faces and recognizes the presence of a weak maxima also. The frame-based approach is unable to do this, as the maxima is below the applied threshold.*

combinations, which preserve the geometric relations among the characteristic structures. This approach enables correct detection results to be obtained in an acceptable computation time. Many advantages of the wavelet based description, make it an interesting approach for further investigation also in the context of interest points. The wavelets must be simultaneously adapted to the image signal, to the quantization and to the coding process, therefore further experiments with different wavelet functions and coding methods will be valuable.

The classification models were trained on a large number of examples but most significant face features were selected manually, that is by looking at the detection results in the training set. This process can be improved by automatic boosting algorithms as i.e. AdaBoost [100]. We can also improve and accelerate the decision process by applying a hierarchical approach [33]. The preliminary decision can be based on more prominent face attributes and only the regions with higher probability would be examined in detail.

The face detection approach was extended and adapted to the detection in video sequences. The use of temporal correlation between parameters estimated in consecutive frames of a video sequence significantly improves the performance of the detector. There are several novel contributions of the temporal approach. The accumulation of probabilities of detection over a sequence obtains a coherent detection over time as well as independence from thresholds. The prediction of the detection parameters, which are position, scale and pose guarantees the accuracy of accumulation as well as a continuous

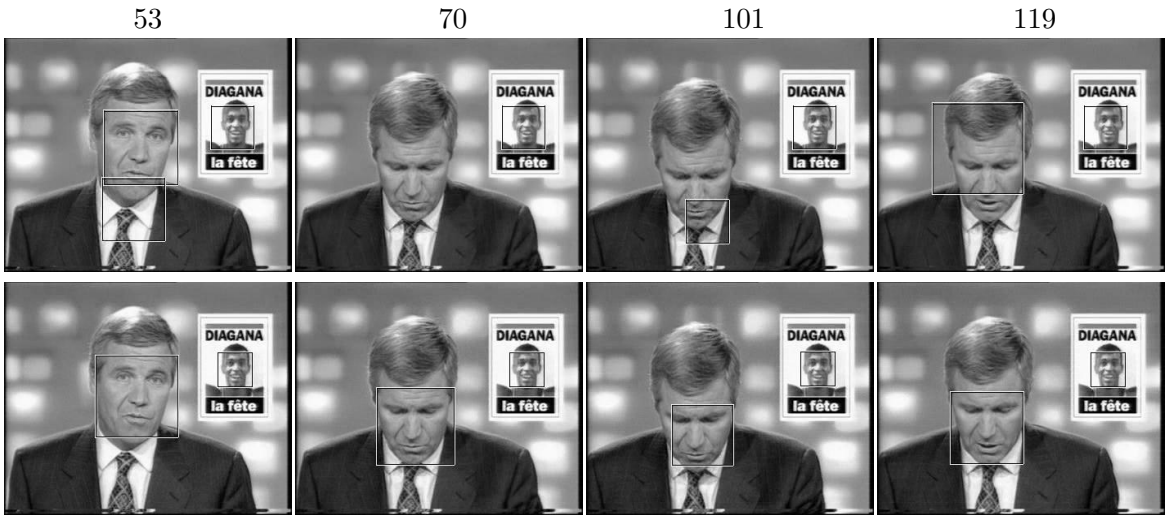


FIG. 7.14: *Detectors applied to a sequence with face nodding up and down. The first row shows faces detected with the frame-by-frame detector. The second row shows results of the "temporal" detector. .The faces continue to be tracked even in the case of a head movement and out-of-plane rotation, while the frame-by-frame detection loses some of these faces as the probability score associated drops below a threshold. The temporal approach is able to handle up-down head movement whereas the detector fails as it has not been trained with such face poses. In frame 70 we see that the bigger face is not detected and in frame 101, is incorrectly detected. The temporal approach improves these results.*

detection. The representation of pose is based on the combination of two detectors, one for frontal views and one for profiles. Thus, the actual pose of the face is estimated for each frame using a combination of the responses from these two detectors. The proposed approach presents several advantages over the existing approaches. Incorporating the temporal information significantly reduces the search area. Our approach is also able to handle changes in imaging conditions (face scale, shadows, lighting and orientation) and changes in image content (the complexity of the background and the number of faces). Furthermore, the proposed temporal model is more robust as it inherently takes into account the variation over time (the detector has already incorporated these variations over a large set of examples), as opposed to traditional tracking approaches which learn the temporal variation over time. Also compared to existing detection and tracking methods, we avoid initialization problems. The ability of our framework to handle pose makes it possible to deal with out of plane rotations, which is considered to be a challenging task for any tracker.

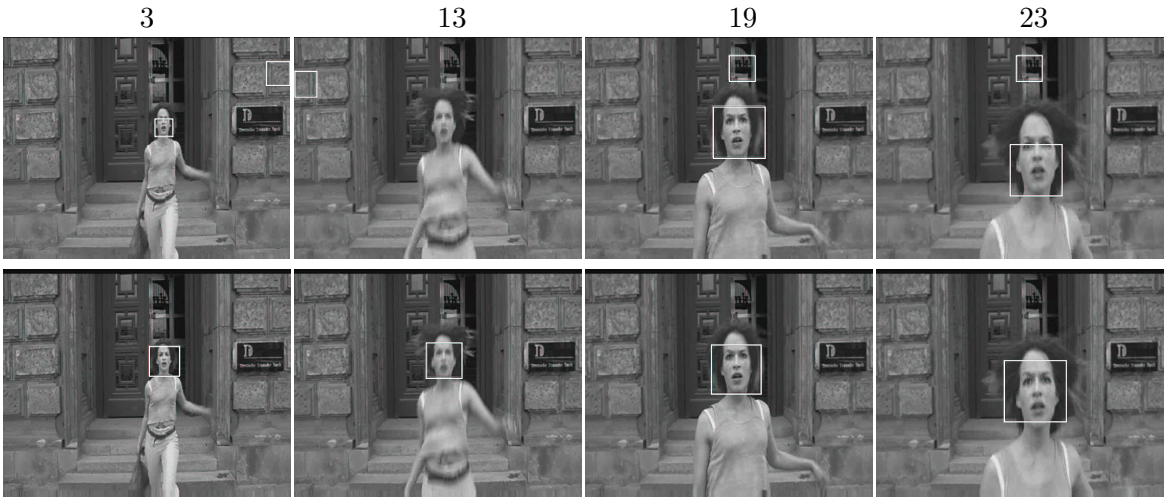


FIG. 7.15: *Detectors applied to a zoom sequence. The first row shows faces detected with the frame-by-frame detector. The second row shows results of the "temporal" detector. In frame 3, 19 and 23, there are false detections, which are eliminated by the temporal approach. Unlike the temporal approach the detector is unable to find the face in frame 13.*

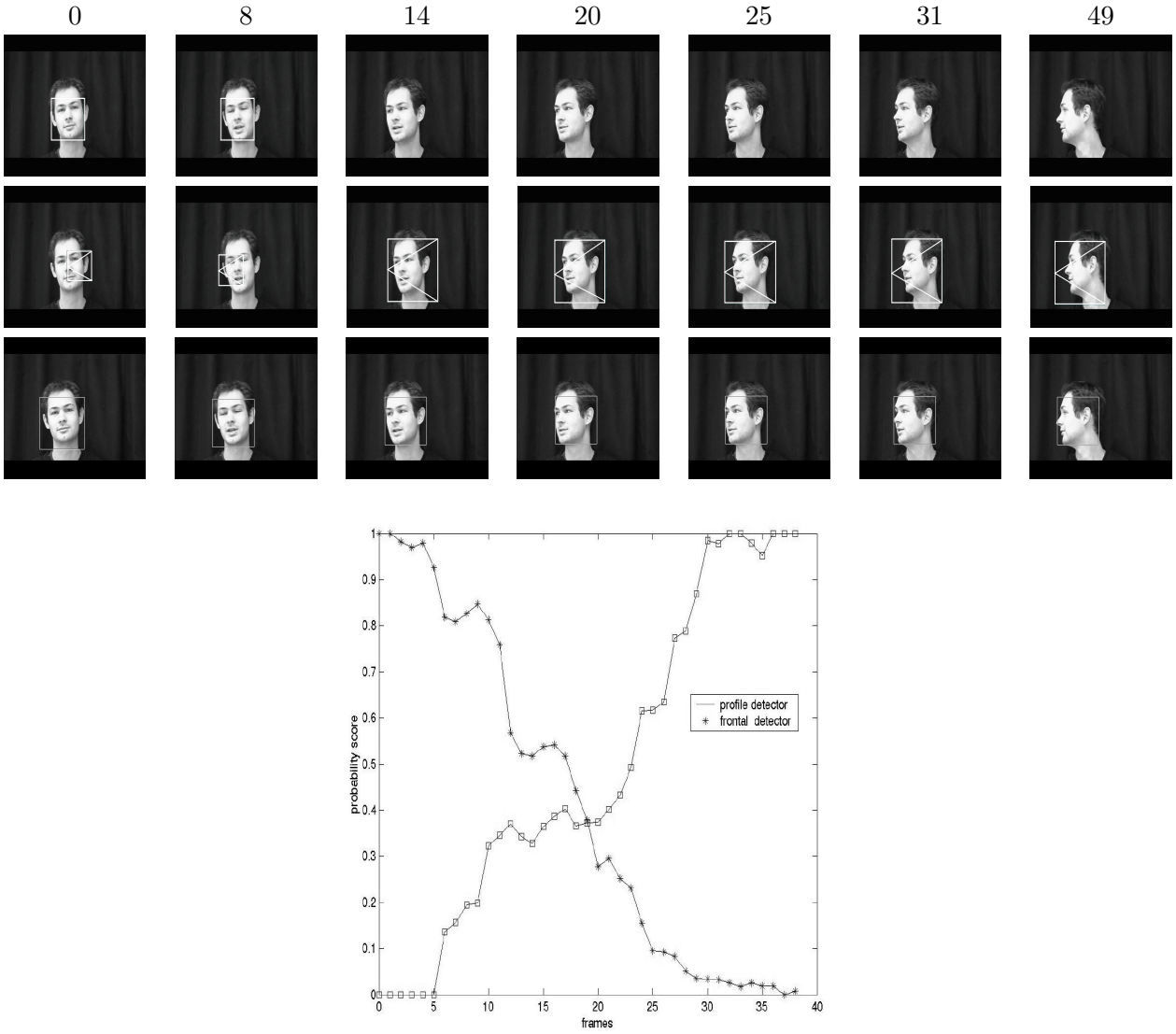


FIG. 7.16: *Variation of head pose. The responses are represented by two different types of boxes, square for the frontal and a square embedded with a triangle for the profile. The first row shows the results of applying only the frontal detector to the sequence. The face is not detected when its left part disappears due to rotation of the head. The second row shows the results of applying the profile detector. In this case the frontal faces are incorrectly detected. This indicates that a single detector is insufficient. In the temporal approach, the responses from the two detectors are combined and the face is continuously tracked through the sequence. The third row shows the results of applying the temporal approach. The face is detected in all frames.*

Chapter 8

Discussion

A novel approach for detecting interest points was proposed in this thesis. The main contributions of our work are the feature detectors, which are invariant to significant geometric transformations, frequently present in images. In the following sections we present the conclusions and the opportunities for future work.

8.1 Conclusions

Our approach provides for reliable recognition in the presence of significant changes in viewing conditions. The detection of local characteristic points is the first step in the process of matching or recognition. The robustness and invariance of these features is therefore crucial for these applications.

We have proposed two novel approaches for the detection of invariant interest points, which are the main contributions of the thesis. The algorithms are based on strong mathematical foundations, the usefulness of which has been confirmed by numerous research results in the feature detection domain. The two approaches are related by the iterative algorithm, which enables the localization and the scale of points to be found with high accuracy. The extended version additionally estimates the affine transformation of the point neighborhood.

The approach proposed in this dissertation provides for reliable detection truly invariant to affine transformations. It requires no prior knowledge of the transformation between images. All the parameters affected by an affine transformation, that is the localization, the scale, the rotation and the shape of the local structure are estimated in an affine invariant way. It is the first approach that can handle simultaneously affine photometric and geometric transformations including significant scale changes. This invariance is not due to the descriptors, therefore any description technique can be used to represent the features and the geometric invariance will be preserved.

We have also proposed a simplified scale invariant algorithm designed to handle the frequent problem of scale change between images taken from different distances or with

different focal settings. This approach is based on two detectors, Harris and Laplacian, both of which have been previously presented in literature, but separately. The combination of these two methods achieves invariance to significant scale changes, which presented a strong limitation in previous approaches to image recognition. The iterative method proposed in this algorithm can be used to find the exact location and scale of a local structure.

The method based on local extrema in the scale-space representation is currently the most reliable and efficient scale selection approach. It is important to select the best method for each part of the feature detection algorithm in order to find as many potentially stable features as possible. We have carried out an experimental evaluation of automatic scale selection, which proved the usefulness of this method. Our comparative test identified the scale selection operator, which gives the best results. We have compared our approach to the methods presented in literature. To obtain representative results we have established evaluation criteria, which take into account the essential parameters affected by the considered transformations. This evaluation is done on real images and shows the excellent repeatability and accuracy of our interest point detectors. Our method gives better results than any of the existing approaches.

The scale and the affine invariant Harris detector was applied to the retrieval of images with large changes in scale and in viewing angle. The results obtained for a database of 5000 images confirm the usefulness of our detectors in this context. The scale invariant detector allows for larger scale changes but fails in the case of wide viewpoint changes. The affine invariant approach gives reliable results in the second case, but a more robust descriptor should be applied to make it robust to large scale changes.

In this thesis we have also proposed and evaluated a stable method for estimating the dominant orientation in a local neighborhood. It can be used to rotate the local structure to compensate for an arbitrary rotation. Hence, any descriptor can be applied to represent the local image structure and the rotation invariance will be preserved.

In the domain of local description our contribution is the analysis, which identifies the drawbacks of the differential descriptors. They are sensitive to noise and to the accuracy of interest point location. Therefore, more robust descriptors are required. We have carried out a preliminary evaluation of differential descriptors with criteria based on the entropy measure. We have presented the comparative results for two types of differential descriptors. The results are easy to interpret and confirm our observations that the steerable filters are more distinctive than the differential invariants. In future work, we can compare a larger set of descriptors using these criteria.

The second part of the thesis concerns the detection of an object class, the human face. We are confronted with the problems of recognition based on local features without using 3D models. Presently, we are able to detect the human face in the presence of partial occlusions, scale and illumination changes. The approach does not use color, it is therefore robust to color changes, frequently present in professional video. The face detector is based on the distributions of local characteristics of faces, accumulated in a histogram. This approach represents a compromise between the methods, which are efficient in computation time but not robust, and the methods, which give reliable results, but are very time consuming.

We have implemented an approach based on the wavelet transform. Excellent detection results prove the usefulness of the descriptors based on wavelet coefficients. The combinations of wavelet coefficients properly capture the essential attributes of the face. Moreover, the wavelet representation is very compact, therefore we can use many different descriptors. The combination of coefficients from different frequency bands and orientations preserve the geometric relations among the characteristic structures, which is a very important property of object description. The accumulated probability of appearance of different face attributes provides for a robust classification in the presence of shadows and occlusions.

We have proposed a method to determine the pose of the face. We use two independent face models, one for front views and the other for profiles, in order to detect faces, which can appear in any pose. The classification models are trained on a set of examples, which contain faces in intermediate poses. This provides the positive responses from both detectors for the intermediate pose. Furthermore, we have shown that these two responses can be related by a model, which enables the actual face pose to be predicted.

We have introduced a novel approach for temporal face detection, which significantly improves detection in single video frames. The detection parameters: location, scale, pose and probability of face appearance are predicted and updated with the Condensation filter. This improves the detection results. The temporal information eliminates the trembling effect of the bounding box, which is due to slightly different location detected in consecutive frames. The false positives are removed as they are rarely present in several consecutive frames, and therefore easy to identify. The spurious detections are solved with the results provided by the neighboring frames. The temporal approach also improves the robustness against shadows and partial occlusions. It also reduces the search space and therefore significantly accelerates the computation.

The excellent results, which we have obtained with the approaches proposed in this manuscript, lead us to believe that the algorithms will be influential on future work and will contribute to further progress in related domains.

8.2 Future work

The opportunities arising from this work can be divided into two complementary parts. The first involves matching of wide baseline images and recognition of objects, which are viewed in different conditions. The second one concerns the classification of similar objects. The matching of images using interest points was developed a few years ago, and since then many solutions have been proposed, which give satisfying results. On the other hand there are many unsolved problems in the domain of recognition and classification of objects and new successful approaches are still required. Characteristic features useful for recognition are different for various objects, for some objects they may be edges for others corners or blobs. An algorithm could automatically choose the appropriate features given in a set of examples. Other problems are related to the process of training a model and capturing the statistical distribution of features. In the following we briefly present some of our ideas for further research in the recognition domain.

8.2.1 Matching and recognition of rigid objects

Descriptors. In the experimental evaluation we have seen that the performance of the matching approach is limited by the differential descriptors, which are sensitive to noise and not sufficiently distinctive. Therefore, a new robust and distinctive descriptor is required. We can improve the description by combining different types of information, in particular, the information represented by the distribution of color or texture. In our approach we do not use the information represented by color. The descriptors, which use **color** can certainly better represent a local image structure. However color information is very sensitive to illumination changes. Shadows, the color of the light as well as the direction of the light source can significantly change the information conveyed by color. The global approaches based on histograms appeared to be inefficient, insufficient and difficult to adapt to the local approach. Color based descriptors have to be invariant to illumination conditions and still convey essential information. To obtain such descriptors an extensive study of color representations in the context of local description is required. There are several ways of representing the color. The representations RGB, Luv, Lab, HSV enable different aspects of the information to be analyzed. There are also different models for color change under lighting conditions. These models have to be evaluated in order to select the most representative one. The diagonal model seems to be a compromise between the accurate but more complex and the simple but less representative models [44]. The separation of luminance from chrominance in the color opponent space used for color invariants [60] provides some robustness to illumination changes when the luminance axis is normalized for energy. Ter Haar Romeny [121] proposes the normalization of color invariants to intensity gradients. Only color gradients are detected by such an operator. To obtain the invariance, we can either compute the invariant descriptor or normalize the color of the image structure, on which the descriptors are computed. The choice between these two solutions is not easy and experimental evaluation of each technique is necessary.

Texture is also a distinctive property of local features. It can be, for example, represented by a distribution of repeated motifs. We can also decompose the signal into basis frequencies using Gabor filters or the wavelet transform. In the first case we have to select the filters, which capture the changes in frequency and orientation. In the second case we have to select and combine the wavelet coefficients to represent the essential information. A comparative study of description techniques, presented in the literature, would be valuable. It will determine robust and distinctive descriptors.

Appearance model. A model accumulating the characteristics extracted from multiple views can reliably represent the appearance of an object [74, 98]. Such a model can be a space of features that preserves the internal relations among them. Hence, we are provided with a complete model and we can apply an additional verification using the localization, the scale and the orientation related to the object. This approach can also be used to classify the object views. The final verification of correct recognition has to involve all the parameters affecting the probability of a match, as for example the size of the database, the number of features representing the object, the similarity among the descriptors and the pose constraints, if possible. Different types of image structures and the combination of these structures [61] will enrich the local description of images. Recently many

new extraction techniques appeared in literature. Hence, we can reliably extract interest points, blobs and edges invariant to scale and affine transformations. We can use all these primitives and combine them to describe the relations among them.

Probabilistic approach. To obtain reliable matching results we can apply the approaches based on robust statistics [105]. The probability of a correct match can be associated to the number of potential matches and to the similarity distance. The distinctive character of the descriptor depends on the number of features, which are similar in terms of the similarity measure. The higher the number the lower the distinctiveness. The total number of descriptors in the database can also affect their distinctiveness. Given a transformation between two corresponding points we can increase the probability of a correct match if the neighboring matches follow the transformation. Such a probabilistic matching process can be done in the Bayesian framework.

Temporal information. In the context of a video sequence we can use the information redundancy of consecutive images to increase the robustness of descriptors. This can be done by using only the features present in several images. Another possibility is to compute and to model the variability of each descriptor with a statistical distribution. This involves the analysis of different distributions in order to find the most representative one for descriptors. The criteria for model selection have to be adapted to the problems related to interest points. We have to consider the high dimensionality of descriptors and the correlation between the components. In the case of video it can be a dynamic model, updated after each correct recognition. An accumulation of detection parameters in the sequence provides a coherent detection in the time, and simultaneously tracks the object. A prediction of the parameters reduces the search area and accelerates the process. The temporal approach can be used for matching video sequences, synchronizing the sequences, which are filmed from different viewpoints and reconstructing dynamic 3D scenes.

Applications. One future area of work consists in validating the proposed approaches in different applications. There are many applications to which our approach can be applied (cf. section 2.1). We have proposed a general solution for feature detection, but it may be useful to adapt this approach to specific applications such as for example the analysis of a video sequence. Similarly, as was done for the face detector, we can propagate the essential parameters in time and therefore reduce the search space. This will certainly accelerate the iterative algorithm, and enable the stable features, which are present in the consecutive images, to be identified.

8.2.2 Recognition of an object class

The classification of objects is a relatively new subject in computer vision and the proposed solutions need to be improved. New approaches are still required, as the existing ones do not provide satisfying results or are adapted to specific problems.

Descriptors. In order to obtain a classification, which is robust to occlusions and background clutter we can use methods based on local descriptors combined with statistical approaches. The attributes of an object are present at different scale levels. Multi-resolution representations are usually explored to extract local features invariant to geometric transformations. One of such representations, frequently used in image compression algorithms,

is obtained with the wavelet transform. The property of compact representation is particularly useful when the appearance of the object is very complex and requires an extensive description. The image structures can be decomposed into the basis frequencies and combined with respect to frequency, orientation and scale. Thus, they are more robust against different transformations caused by a change in viewing conditions. As we have seen in chapter 7 the wavelet transform can handle all these problems. Furthermore, to enrich the description we can take into account the locations of local features relative to the internal object coordinates. The spatial relations between the attributes can also be determined in order to obtain more distinctive description. We can also incorporate additional cues like color or texture.

Classification models. The process of classification is usually done between two classes. One class represents the appearance of an object in different conditions and in different instances. The other one represents the features which do not belong to the object. We use positive and negative examples, which represent the two classes. While the selection of positive examples is rather straightforward, that is the images represent the object, the selection of non-object examples is not easy. Statistical methods can be used to construct models. AdaBoost technique [100] can be applied to attribute weights to more significant characteristics. The objective is to attribute the weights, which provide for better recognition results for a given training set of images. Another solution is to use Support Vector Machines [20] to obtain the optimal partition of descriptor space of positive and negative examples. Thus, the models are based on the features which are most representative for each of the class. The face detector, which is described in this manuscript, can also be applied for detecting other complex objects, like cars or pedestrians. This may require different descriptors, as the characteristic attributes of these objects are different.

Decision process. We can also use the hierarchical approach in the classification process [33]. The decision process can be divided into several steps. The preliminary decisions can be based on more significant and distinctive features and only the similar regions would be examined in detail. This accelerates the classification process and increases the number of representative descriptors.

The query image and the models are compared using a similarity measure. This measure should be independent of the thresholds that can occur at different steps of the classification. The threshold based decisions can be replaced by an accumulated probability updated at each step of the classification process.

Temporal information. In the context of video sequence we can use the temporal information to obtain more robust and reliable recognition. The detection and tracking information can be integrated in the classification model. We can apply a dynamic model, which can be updated after each correct classification. The updating is more valuable, when the query object is classified with lower certitude. Unfortunately, this involves additional methods for verifying the correctness of the classification, otherwise the inappropriate model may be updated. Usually, the verification is not applied at each video frame, therefore more complicated approaches can be used, which are usually more time consuming. Such approaches are prohibitive in the context of video. The updated model can be used for dynamic object recognition, i.e. for human faces. To reduce the search area we can also incorporate a motion model. For this purpose the Kalman filter [41] or the Condensation

algorithm [52] can be used. In future work we could apply a motion model to keep track on occluded faces.

The detection, recognition and classification algorithms, all have to be validated in the context of real applications. That is the only way to determine the advantages and the drawbacks of an approach.

Annexe A

ANNEX

A.1 Extremum of local derivatives

In the following we derive a relation between the scale parameter of the second order Gaussian derivative and the maximum response of convolution with a step-edge function. The necessary condition to find a local extremum of a function F_{norm} over scale is $\frac{\partial}{\partial\sigma}F_{norm} = 0$, where σ is the scale factor. Given a function representing the step-edge presented in figure 3.8(c),(f):

$$f_{step-edge}(x,y) = \begin{cases} 0 & x < x_0 \ \forall y \\ 1 & x \geq x_0 \ \forall y \end{cases}$$

we can compute the normalized second derivatives in point $(0,0)$:

$$f_{xx_{norm}}(x,\sigma) = g_{xx}(\sigma) * f_{step-edge}(x,y) =$$

$$f_{xx_{norm}}(x,\sigma) = \sigma^2 \int_{-\infty}^{+\infty} g_{xx}(n,\sigma) f(x-n) dn \int_{-\infty}^{+\infty} g(m,\sigma) f(y-m) dm$$

the function $f_{step-edge}$ is constant in the y dimension therefore:

$$\int_{-\infty}^{+\infty} f(y-m,x) g(m,\sigma) dm = f_{step-edge}(x)$$

$$f_{xx_{norm}}(x,\sigma) = -\sigma^2(g_x(x_0,\sigma)) = \frac{x}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

In order to find an extremum over scale we compute $\frac{\partial}{\partial\sigma} f_{xx_{norm}}(x,\sigma) = 0$.

$$\frac{\partial}{\partial\sigma} \frac{x}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} (-\frac{x}{\sigma^2} + \frac{x^3}{\sigma^4}) = 0 \Leftrightarrow \sigma_{extremum} = |x_0| \tag{A.1}$$

The second order derivative attains an extremum when the kernel size is equal to the distance from the signal change $|x_0|$.

A.2 Repeatability criterion

In this section we explain in detail how we compute the intersection between two regions determined by affine covariant interest points. The following equations are used in section 4.3 to evaluate affine invariant interest point detector. The error in image surface ϵ_S covered by point neighborhoods is:

$$\epsilon_S = 1 - \frac{\mu_a \cap (A^T \mu_b A)}{(\mu_a \cup A^T \mu_b A)} < 0.2$$

where μ_a and μ_b are the elliptic regions defined by $x^T \mu x = 1$. The union of the regions is $(\mu_a \cup (A^T \mu_b A))$ and $(\mu_a \cap (A^T \mu_b A))$ is their intersection. A is a locally linearized homography H in point $\mathbf{x}_b$. To compute ϵ_S we transform the ellipse μ_b to the coordinate

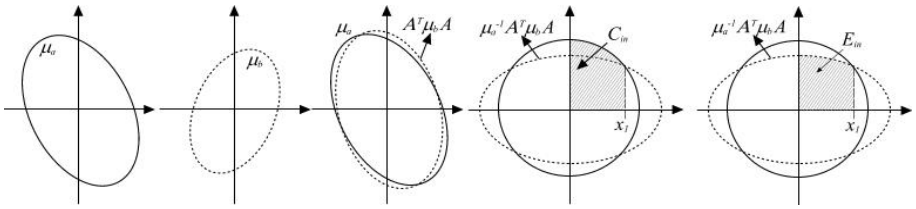


FIG. A.1: *Surface intersection of elliptical affine points.*

system of the ellipse μ_a, that is $\mu_b' = A^T \mu_b A$. Next we apply a transformation that brings the ellipse μ_a to a circle with radius $r = 1$, and ellipse μ_b' to $\mu_a^{-1} A^T \mu_b A$. The eigenvalues of the ellipse are λ_{max} and λ_{min}. Note that the surface ratio is preserved under an affine transformation and ϵ_S do not change if we apply the same linear transformation to both regions. The intersection point x_1 is the solution of:

$$\sqrt{r^2 - x^2} = \sqrt{\lambda_{min}^2 - \frac{\lambda_{min}^2}{\lambda_{max}^2} x^2}$$

If there is no intersection between the circle and the ellips the surface error is:

$$\epsilon_S = \frac{|\Pi r^2 - \Pi \lambda_{max} \lambda_{min}|}{min(\Pi r^2, \Pi \lambda_{max} \lambda_{min})}$$

otherwise the area C_{out} that belongs to the circle and not to the ellipse is:

$$C_{out} = 4[\int_0^{x_1} \sqrt{r^2 - x^2} dx - E_{in}]$$

$$E_{in} = \int_0^{x_1} \sqrt{\lambda_{min}^2 - \frac{\lambda_{min}^2}{\lambda_{max}^2} x^2} dx$$

the intersection of the regions is:

$$C_{in} = \Pi r^2 - C_{out}$$

$$E_{in} = \int_0^{x_1} \sqrt{\lambda_{min}^2 - \frac{\lambda_{min}^2}{\lambda_{max}^2} x^2} dx = \frac{\lambda_{min}}{\lambda_{max}} \int_0^{x_1} \sqrt{\lambda_{max}^2 - x^2} dx =$$

$$x = \lambda_{max} \sin(t),\ dx = \lambda_{max} \cos(t) dt,\ t_E = \arcsin(x_1/\lambda_{max})\ t_C = \arcsin(x_1/r)$$

$$E_{in} = \lambda_{min}\lambda_{max} \int_0^{t_E} \cos^2(t) dt = \lambda_{min}\lambda_{max}(\frac{t_E}{2} + \frac{1}{4}\sin(2t_E))$$

$$C_{in} = r^2 \int_0^{t_C} \cos^2(t) dt = r^2(\frac{t_C}{2} + \frac{1}{4}\sin(2t_C))$$

the outside of circle regions is:

$$C_{out} = 4(C_{in} - E_{in})$$

Thus we obtain the intersection error, that is the surface, which is not covered by both regions devided by the intersection of regions:

$$\epsilon_S = \frac{\Pi\lambda_{max}\lambda_{min} - (\Pi r^2 - C_{out}) + C_{out}}{(\Pi r^2 - C_{out})}$$

A.3 Test images

In the following we display the image sequences with scale changes and with perspective transformations, which are used in our matching and recognition experiments in sections 6.2, 6.3, 5.3 and 4.3. The images are also accessible via the Internet:
`http:\\www.inrialpes.fr\movi\people\Mikolajczyk\Database`

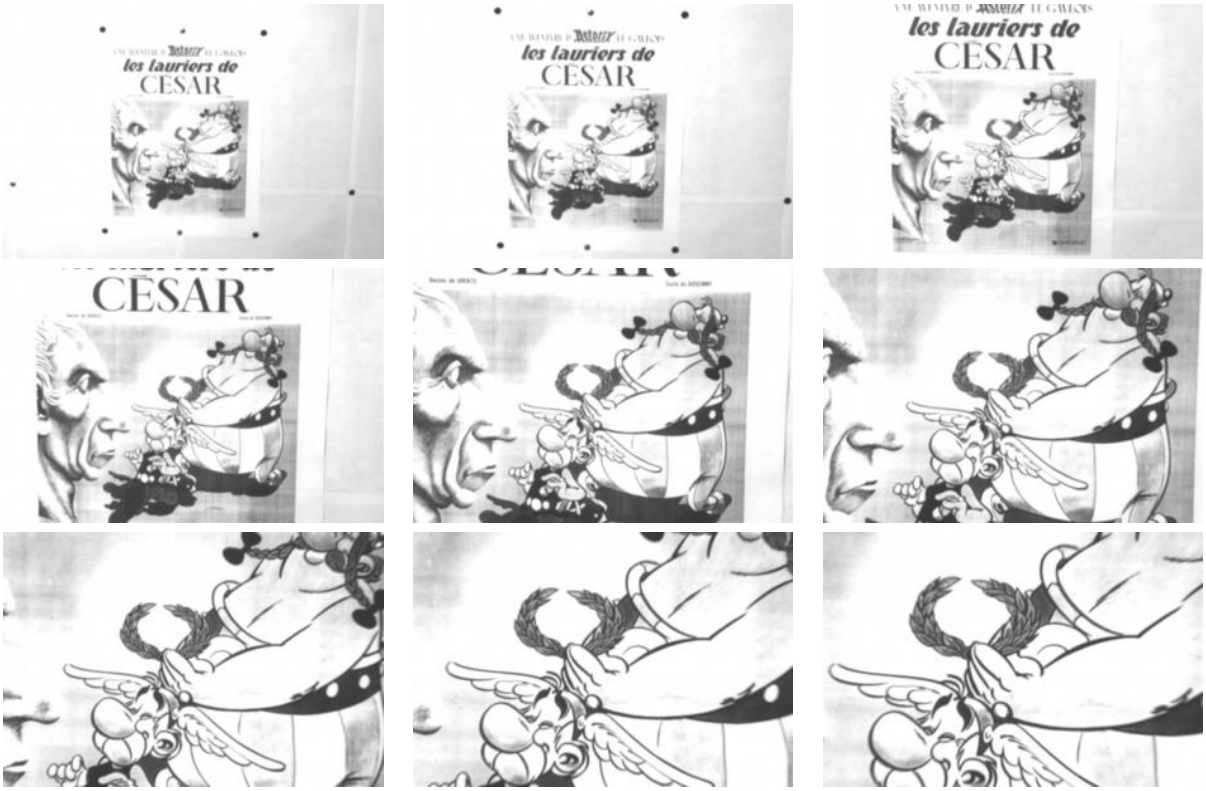


FIG. A.2: *Scene 1: Scale change.*

FIG. A.3: *Scene 2: Scale change.*

FIG. A.4: *Scene 3: Scale change.*

FIG. A.5: *Scene 4: Scale change.*

FIG. A.6: *Scene 5: Scale change.*

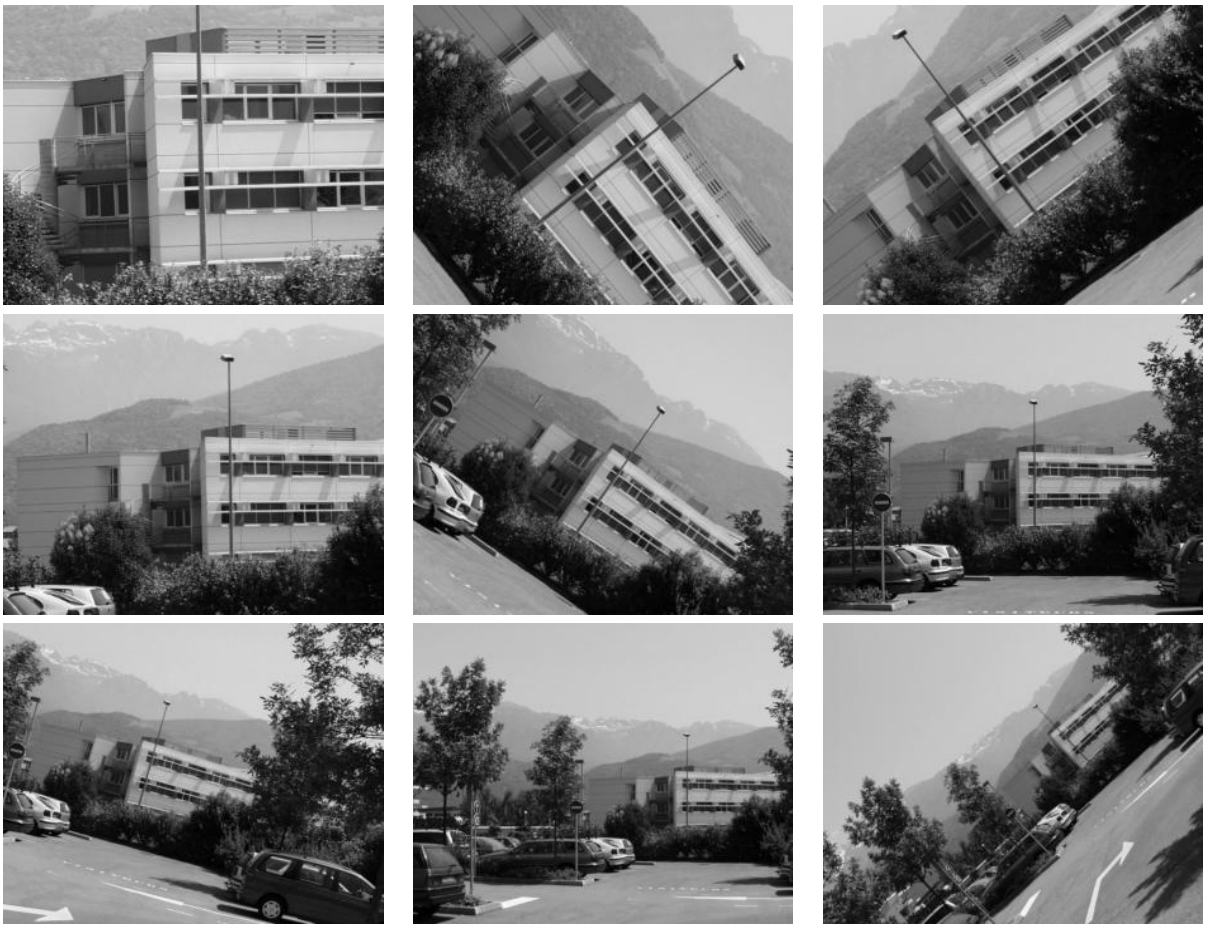


FIG. A.7: *Scene 6: Scale change.*

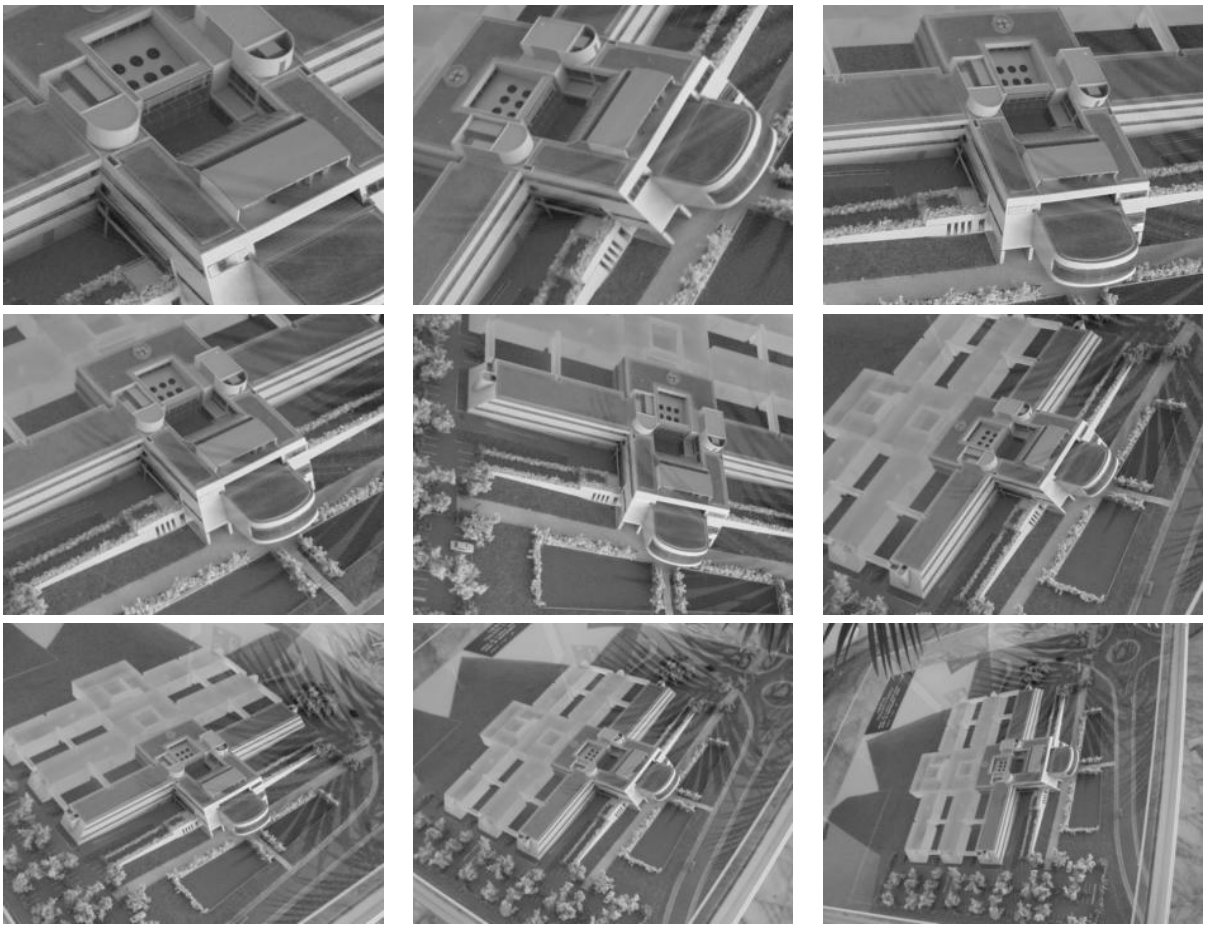


FIG. A.8: *Scene 7: Scale change.*

FIG. A.9: *Scene 8: Scale change.*

FIG. A.10: *Scene 9: Scale change.*

FIG. A.11: *Scene 10: Scale change.*

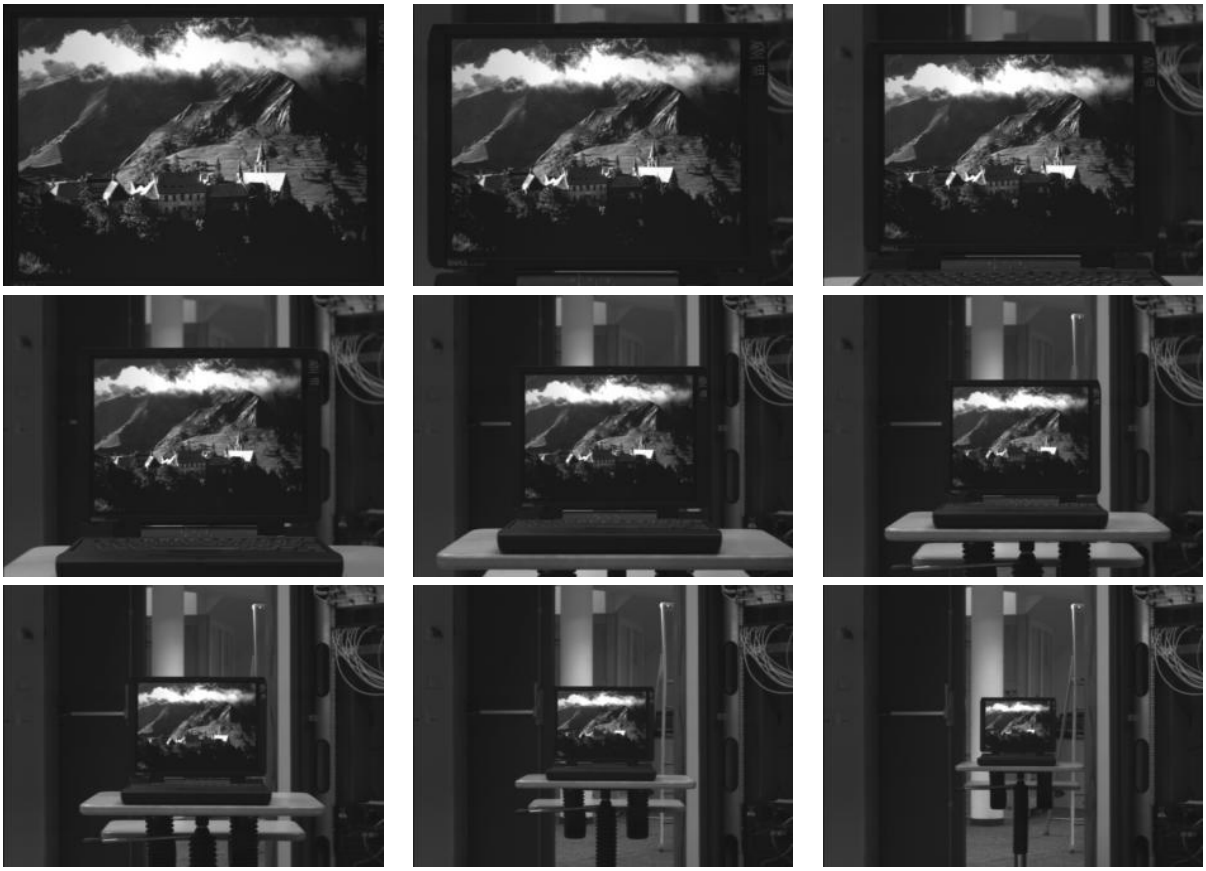


FIG. A.12: *Scene 11: Scale change.*

FIG. A.13: *Scene 12: Perspective transformation.*

FIG. A.14: *Scene 13: Perspective transformation.*

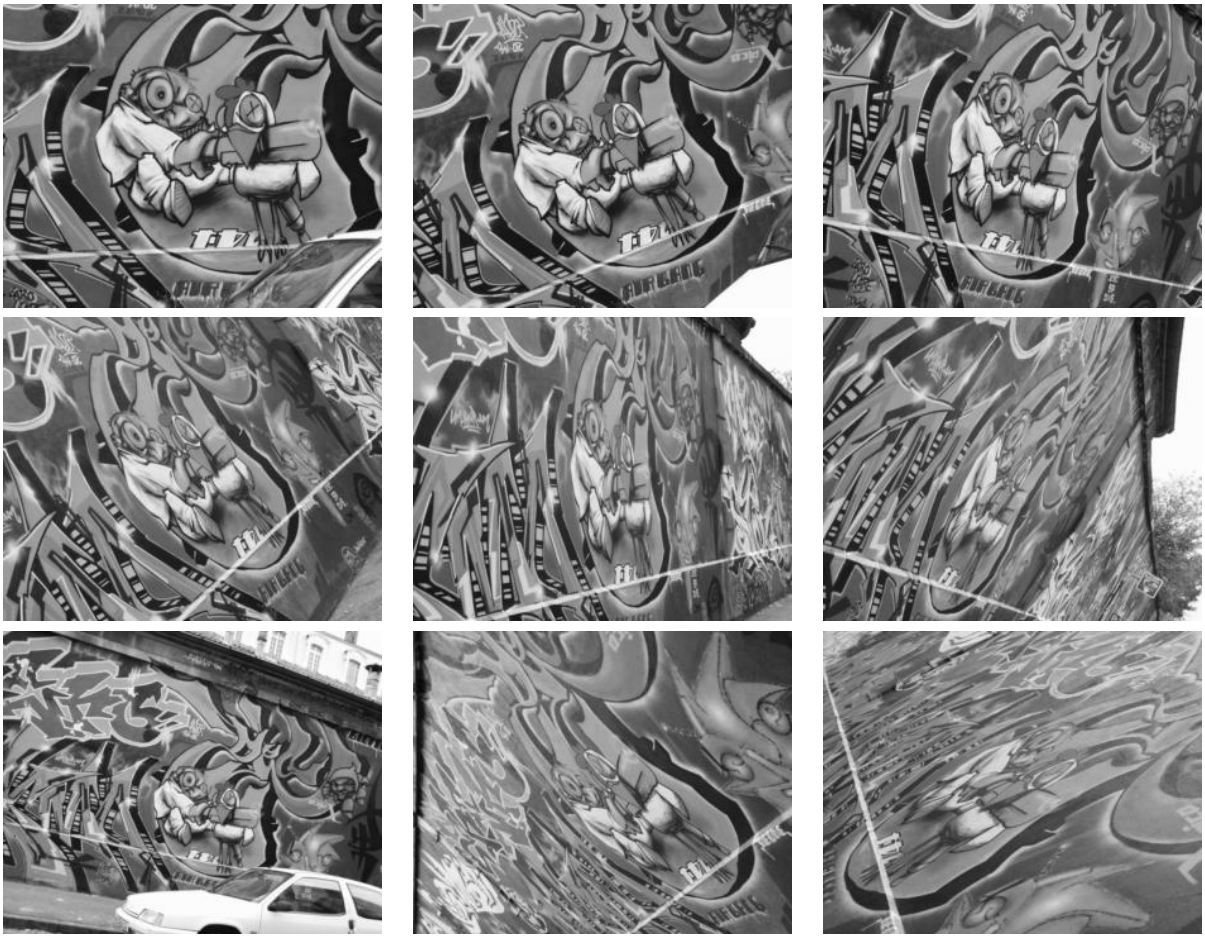


FIG. A.15: *Scene 14: Perspective transformation.*

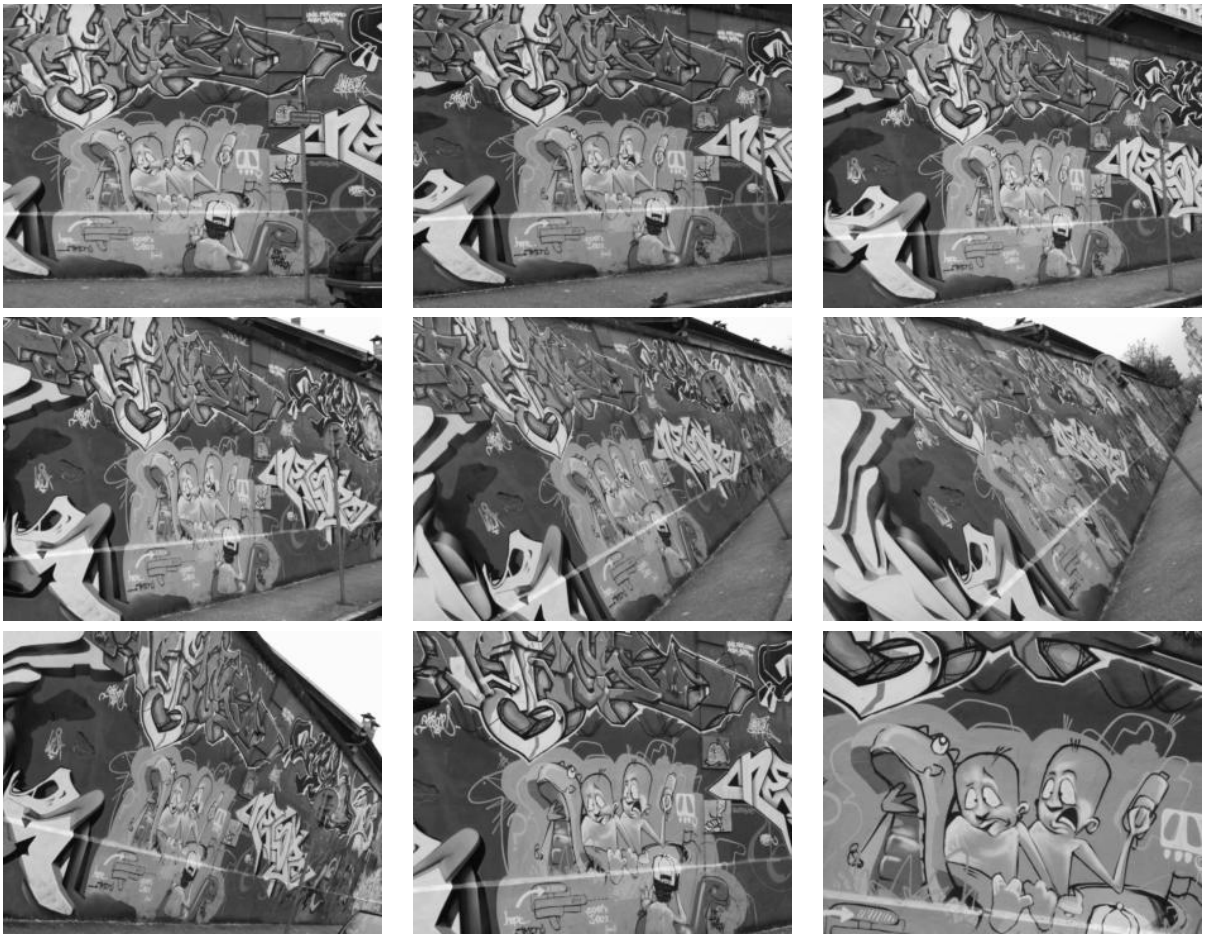


FIG. A.16: *Scene 15: Perspective transformation.*

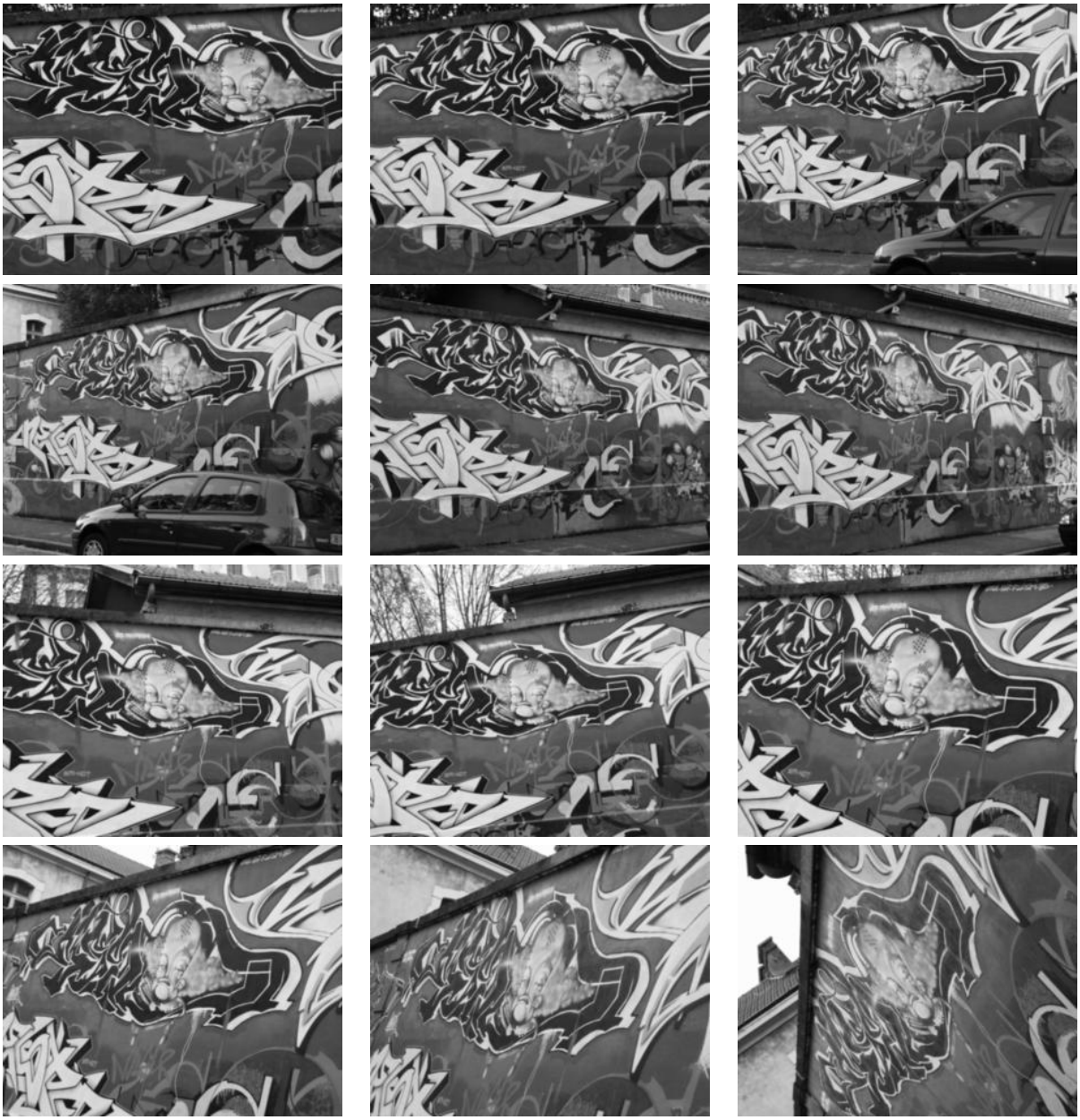


FIG. A.17: *Scene 16: Perspective transformation.*

FIG. A.18: *Scene 17: Perspective transformation.*

References

[1] E.H. Adelson, E. Simoncelli, and R. Hingorani. Orthogonal pyramid transforms for image coding. *SPIE Visual Communication and Image Processing II*, 845:50–58, 1984.

[2] A. Almansa and T. Lindeberg. Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale selection. IEEE *Transactions on Image Processing*, 9(12):2027–2042, 2000.

[3] L. Alvarez and F. Morales. Affine morphological multiscale analysis of corners and multiple junctions. *International Journal of Computer Vision*, 2(25):95–107, 1997.

[4] L. Amsaleg and P. Gros. Content-based retrieval using local descriptors: problems and issues from a database perspective. *Pattern Analysis and Applications*, 2/3(4):108–124, April 2001.

[5] J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda. Uniqueness of the gaussian kernel for scale-space filtering. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 8(1):26–33, 1986.

[6] M. Barlaud, editor. *Wavelets in image communication*. Elsevier, 1994.

[7] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 774–781, 2000.

[8] P.R. Beaudet. Rotationally invariant image operators. In *Proceedings of the 4th International Joint Conference on Pattern Recognition, Tokyo*, pages 579–583, 1978.

[9] J. Bigun and J. M. H. du Buf. Texture segmentation by real and complex moments of the gabor power spectrum. *Progress in Image Analysis and Processing*, II:191–198, 1992.

[10] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232–237, 1998.

[11] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 1073–1080, January 1998.

[12] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using parameter models of image motion. In *ICCV*, pages 374–381, 1995.

[13] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spacial filters. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 12:55–73, 1990.

[14] L. Bretzner and T. Lindeberg. Feature tracking with automatic selection of spatial scales. *Computer Vision and Image Understanding*, 71(3):385–392, 1998.

[15] M. C. Burl, T. K. Leung, and P. Perona. Face localisation via shape statistics. *FG*, pages 154–159, 1995.

[16] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. IEEE *Transactions on Communications*, 9(4):532–540, July 1983.

[17] O. Chomat, V. Colin de Verdière, D. Hall, and J. Crowley. Local scale selection for gaussian based description techniques. In *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, pages 117–133, 2000.

[18] D. Comaniciu, V. Ramesh, and P. Meer. Real time tracking of non-rigid objects using mean shift. In *CVPR*, pages 142–149, 2000.

[19] J.C. Cottier. Extraction et appariements robustes des points d'intérêt de deux images non etalonnées, September 1994.

[20] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

[21] J.L. Crowley. *A Representation for Visual Information*. PhD thesis, Carnegie Mellon University, November 1981.

[22] J.L. Crowley and A.C. Parker. A representation for shape based on peaks and ridges in the difference of low pass transform. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 6(2):156–170, March 1984.

[23] J.L. Crowley and A.C. Sanderson. Multiple resolution representation and probabilistic matching of 2D gray-scale shape. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 9(1):113–121, 1987.

[24] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure & Appl. Math*, 41:909–996, 1988.

[25] D. Decarlo and D. Metaxas. Deformable model based face shape and motion estimation. In *IEEE Proceedings of ICFG*, 1996.

[26] R. Deriche. Recursively implementing the Gaussian and its derivatives. Technical report, INRIA, April 1993.

[27] R. Deriche and G. Giraudon. A computational approach for corner and vertex detection. *International Journal of Computer Vision*, 10(2):101–124, 1993.

[28] J. M. H. du Buf. Abstract processes in texture discrimination. *Spatial Vision*, 6(3):221–242, 1992.

[29] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, 1973.

[30] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 612–618, June 2000.

[31] C. J. Edward, C. J. Taylor, and T. F. cootes. Learning to identify and track faces in an image sequence. In *International Conference on Automatic Face and Gesture Recognition*, pages 260–265, 1998.

[32] R. S. Feris, T. E. de Campos, and R. M. Cesar Junior. Detection and tracking of facial features in video sequences. In *Lecture notes in AI*, pages 197–206, 2000.

[33] F. Fleuret and D. Geman. Apprentissage hierarchique pour la detection de visages. In *12ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, February 2000.

[34] L. Florack. *The Syntactical Structure of Scalar Images*. PhD thesis, Universiteit Utrecht, November 1993.

[35] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever. Scale and the differential structure of images. *Image and Vision Computing*, 10:376–388, 1992.

[36] W. Förstner. A framework for low level feature extraction. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, pages 383–394, 1994.

[37] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Intercommission Conference on Fast Processing of Photogrammetric Data, Interlaken, Switzerland*, pages 281–305, June 1987.

[38] W.T. Freeman and E.H. Adelson. The design abd use of steerable filters. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[39] D. Gabor. Theory of communication. *Journal I.E.E.*, 3(93):429 – 457, 1946.

[40] J. Garding and T. Lindeberg. Direct estimation of local surface shape in a fixating binocular vision system. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, pages 365–376, 1994.

[41] A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, 1992.

[42] J.-M. Geusebroek, A.W.M. Smeulders, and J. van de Weijer. Fast anisotropic gauss filtering. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 99–112, 2002.

[43] S. Gong, S. McKenna, and J. J. Collins. An investigation into face pose distributions. In *International Conference on Automatic Face and Gesture Recognition*, pages 265–270, 1996.

[44] P. Gros. Color illumination models for image matching and indexing. In A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquézar, J.-O. Eklundh, and Y. Aloimonos, editors, *Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain*, volume 3, pages 579–583, September 2000.

[45] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Review*, 15:723–736, 1984.

[46] G. Hager and K. Toyama. X vision : A portable substrate for real-time vision applications. *CVIU*, 69(1):23–37, 1998.

[47] I. Haritaoglu, D. Harwood, and L. davis. W4 :who? when? where? what? a real-time system for detecting and tracking people. In *Automatic Face and Gesture Recognition*, pages 222–227, 1998.

[48] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[49] F. Heitger, L. Rosenthaler, R. von der Heydt, E. Peterhans, and O. Kuebler. Simulation of neural contour mechanism: from simple to end-stopped cells. *Vision Research*, 32(5):963–981, 1992.

[50] R. Horaud, T. Skordas, and F. Veillon. Finding geometric and relational structures in an image. In *Proceedings of the 1st European Conference on Computer Vision, Antibes, France*, pages 374–384, April 1990.

[51] E. Petajan H.P. Graf, T. Chen and E. Cosatto. Locating faces and facial parts. *FG*, pages 41–46, 1995.

[52] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998.

[53] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, December 1991.

[54] A. Johnson and M. Hebert. Object recognition by matching oriented points. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pages 684–689. IEEE Computer Society Press, June 1997.

[55] A.E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[56] K. Kanatani. Geometric information criterion for model selection. *International Journal of Computer Vision*, 26(3):171–189, 1998.

[57] M. Kirby and L. Sirovich. Application of kl procedure for characterization of human faces. *IEEE PAMI*, 12(1):103–108, 1990.

[58] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Pattern Recognition Letters*, 1:95–102, 1982.

[59] J.J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–396, 1984.

[60] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[61] B. Lamiroy and P. Gros. Rapid object indexing and recognition using enhanced geometric hashing. In *Proceedings of the 4th European Conference on Computer Vision, Cambridge, England*, volume 1, pages 59–70, April 1996. Postscript version available by `ftp`[1].

[62] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE PAMI*, 19(7):743–756, 1997.

[63] I. Laptev and T. Lindeberg. Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. In *Proceedings of Scale Space and Morphology Workshop, Vancouver, Canada*, volume 2106, pages 63–74. Lecture Notes in Computer Science, 2001.

[64] E.L. Lehman. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, 1975.

[65] T. Lindeberg. Scale-space for discrete signals. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 12(1):234–254, 1990.

[66] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch - a method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, 1993.

[67] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[68] T. Lindeberg. Direct estimation of affine image deformation using visual front-end operations with automatic scale selection. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 134–141, 1995.

1. ftp://ftp.imag.fr/pub/labo-GRAVIR/MOVI/publications/Lamiroy_eccv96.ps.gz

[69] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[70] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[71] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[72] Z. Liu and Y. Wang. Face detection and tracking in video using dynamic programming. In *ICIP*, 2000.

[73] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 1150–1157, 1999.

[74] D. G. Lowe. Local feature view clustering for 3d object recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, pages 682–688, December 2001.

[75] J. Kovacevic M. Vetterli. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[76] J. MacCormick and Andrew Blake. A probabilistic exclusion principle for tracking multiple objects. In *ICCV*, 1995.

[77] P. Majer. The influence of the γ-parameter on feature detection with automatic scale selection. In *Third International Conference, Scale-Space*, number 2106 in LNCS, pages 245–254, Vancouver, Canada, July 2001. Springer-Verlag.

[78] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[79] B.S. Manjunath, C. Shekhar, and R. Chellappa. A new approach to image feature detection with applications. *Pattern Recognition*, 29(4):627–640, 1996.

[80] D. Marr. *Vision*. W.H. Freeman and Company, San Francisco, California, USA, 1982.

[81] J. Matas, J. Burianek, and J. Kittler. Object recognition using the invariant pixel-set signature. In *The Eleventh British Machine Vision Conference, University of Bristol, UK*, pages 606–615, 2000.

[82] P. Meer, E. S. Baugher, and A. Rosenfeld. Frequency domain analysis and synthesis of image pyramid generating kernels. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 9:512–522, 1987.

[83] Y. Meyer. Principe d'incertitude, bases hilbertiennes, et algebres d'operateurs. In *Bourbaki conference*, volume nr 662, 1986.

[84] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001.

[85] F. Mindru, T. Moons, and L. Van Gool. Comparing intensity transformations and their invariants in the context of color pattern recognition. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 99–112, 2002.

[86] F. Mindru, T. Moons, and L. Van Gool. Recognizing color patterns irrespective of viewpoint and illumination. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA*, pages Vol. 1, 368–373, 1999.

[87] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 786–793, 1995.

[88] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transaction on Pattern Analysis an Machine Intelligencs*, 19(7):696–710, July 1997.

[89] T. Moller, R. Machiraju, K. Muller, and R. Yagel. Evaluation and design of filters using a tylor series expansion. IEEE *Transactions on Vizualization and Computer Graphics*, 3:184–199, 1997.

[90] H. Moravec. Visual mapping by a robot rover. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence, Tokyo, Japan*, pages 598–600, 1979.

[91] J. Morlet. Sampling theory and wave propagation. *NATO ASI Series, Acoustic Signal/Image processing and Recognition*, 1:233–261, 1984.

[92] J.A. Noble. Finding corners. *Image and Vision Computing*, 6(2):121–128, May 1988.

[93] T. Ojala and M. Pietikainen. Unserpvided texture segmentation using feature distributions. *Pattern Recognition*, 32(3):477–486, 1999.

[94] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detecition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pages 130–136, January 1997.

[95] P. Perona. Steerable-scalable kernels for edge detection and junction analysis. In *Proceedings of the 2nd European Conference on Computer Vision, Santa Margherita Ligure, Italy*, pages 3–18, May 1992.

[96] S. Picard. *Etude de descripteur non paramétriques pour l'indexation d'images par le contenu*. Thèse de doctorat, Institut National Polytechnique de Grenoble, GRAVIR – IMAG – INRIA Rhône–Alpes, November 1999.

[97] T. Poggio and D. Beymer. Learning networks for face analysis and synthesis. In *International Conference on Automatic Face and Gesture Recognition*, 1995.

[98] A.R. Pope and D.G. Lowe. Learning object recognition models from images. In *Proceedings of the 4th International Conference on Computer Vision, Berlin, Germany*, pages 296–301, 1993.

[99] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 754–760. IEEE Computer Society Press, January 1998.

[100] Y. Singer R. E. Shapire. Improving bossting algorithm using confidence-rated predictions. *Machine Learning*, 37(3):297–336, December 1999.

[101] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using color. In *International Conference on Automatic Face and Gesture Recognition*, pages 228–233, 1998.

[102] H. A. Rowley, S. Baluja, and T. Kanade. Neural networks based face detection. *IEEE PAMI*, 20(1):22–38, 1998.

[103] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, Vancouver, Canada, 2001.

[104] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002.

[105] C. Schmid. A structured probabilistic model for recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA*, volume II, pages 485–490, 1999.

[106] C. Schmid. Constructing models for content-based image retrieval. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, 2001.

[107] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[108] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 230–235, January 1998.

[109] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[110] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, volume I, pages 746–751, 2000.

[111] K. Schwerdt and J. Crowley. Robust face tracking using colour. In *International Conference on Automatic Face and Gesture Recognition*, pages 90–95, 2000.

[112] A. W. Senior. Recognizing faces in broadcast video. In *RATFG-RTS 99*, pages 105–110, 1999.

[113] S.D. Shapiro. Feature space transforms for curve detection. *Pattern Recognition*, 10(3):129–143, 1978.

[114] A. Shashua. Algebraic functions for recognition. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, August 1995.

[115] J. Sporring, M. Nielsen, L. Florack, and P. Johansen. *Gaussian Scale-Space Theory.* Springer-Verlag, 1997.

[116] M.A. Stricker and M. Orengo. Similarity of color images. In SPIE *Conference on Storage and Retrieval for Image and Video Databases V*, volume SPIE-2420, pages 381–392, 1995.

[117] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[118] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, pages 814–828, 2000.

[119] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 814–828, 2002.

[120] B.M ter Haar Romeny, L.M.J. Florack, A.H. Salden, and M.A. Viergever. Higher order differential structure of images. *Image and Vision Computing*, 12(6):317–325, 1994.

[121] B.M. ter Haar Romeny, J. Geusebroek, P. Van Osta, R. van den Boomgaard, and J. Koenderink. Color differential structure. In *Proceedings of Scale Space and Morphology Workshop, Vancouver, Canada*, volume 2106, pages 353–361. Lecture Notes in Computer Science, 2001.

[122] J-C. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *International Conference on Automatic Face and Gesture Recognition*, pages 54–61, 2000.

[123] V. Torre and T.A. Poggio. On edge detection. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 8(2):147–163, 1986.

[124] K. Toyama and A. Blake. Probabilistic exemplar based tracking in a metric space. In *ICCV*, volume 2, pages 50–57, 2001.

[125] B. Triggs. Joint feature distributions for image correspondence. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages II, 201–208, 2001.

[126] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Maui, Hawaii, USA*, pages 586–591, 1991.

[127] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Int. Conf. on Visual Information Systems*, pages 493–500, 1999.

[128] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *The Eleventh British Machine Vision Conference, University of Bristol, UK*, pages 412–425, 2000.

[129] R. Weber and P. Zezula. A quantitative analysis of performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th VLDB Conf.*, 1998.

[130] D. A. White and R. Jain. Similarity indexing: Algorithms and performance. In *SPIE Storage and Retrieval for Image and Video Database*, pages 62–73, 1996.

[131] A.P. Witkin. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, Germany*, pages 1019–1023, 1983.

[132] H.J. Wolfson. Model-based object recognition by geometric hashing. In O. Faugeras, editor, *Proceedings of the 1st European Conference on Computer Vision, Antibes, - France*, pages 526–536. Springer-Verlag, April 1990.

[133] C. Wren, A. Azerbayejani, T. Darrell, and A. Pentland. Pfinder : a real-time tracking of human body. *IEEE PAMI*, 19(7):780–785, 1997.

[134] X. Wu and B. Bhanu. Gabor wavelets for 3D object recognition. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 537–542, 1995.

[135] J. Yang and A. Waibel. Tracking human faces in real time. Technical Report CMU-CS-95-210, CMU, 1995.

[136] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, pages 151–158. Springer-Verlag, May 1994.