Toward semantic-shape-context-based augmented descriptor

Samir Khoualed

N° d'ordre: 2297
EDSPIC: 585

UNIVERSITÉ BLAISE PASCAL - CLERMONT-FERRAND II

ÉCOLE DOCTORALE
SCIENCES POUR L'INGÉNIEUR DE CLERMONT-FERRAND

Thèse

présentée par

SAMIR KHOUALED

pour obtenir le grade de

DOCTEUR D'UNIVERSITÉ
SPÉCIALITÉ: VISION POUR LA ROBOTIQUE

Toward Semantic-Shape-Context-Based Augmented Descriptor

Descripteurs augmentés basés sur l'information sémantique contextuelle

Soutenue publiquement le 29 novembre 2012 devant le jury:

M. Michel DHOME	Président
M. Serge MIGUET	Rapporteur
M. Marco CRISTANI	Rapporteur
M. Frédéric CHAUSSE	Examinateur
M. Umberto CASTELLANI	Examinateur
M. Thierry CHATEAU	Directeur de thèse

RÉSUMÉ

Les techniques de description des éléments caractéristiques d'une image sont omniprésentes dans de nombreuses applications de vision par ordinateur. Nous proposons à travers ce manuscrit une extension, pour décrire (représenter) et apparier les éléments caractéristiques des images. L'extension proposée consiste en une approche originale pour *apprendre*, ou estimer, la présence *sémantique* des éléments caractéristiques locaux dans les images. L'information sémantique obtenue est ensuite exploitée, en conjonction avec le paradigme de *sac-de-mots*, pour construire un descripteur d'image performant.

Le descripteur résultant, est la combinaison de deux types d'informations, locale et contextuelle-sémantique. L'approche proposée peut être généralisée et adaptée à n'importe quel descripteur local d'image, pour améliorer fortement ses performances spécialement quand l'image est soumise à des conditions d'imagerie contraintes.

La performance de l'approche proposée est évaluée avec des images réelles aussi bien dans les deux domaines, 2D que 3D. Nous avons abordé dans le domaine 2D, un problème lié à l'appariement des éléments caractéristiques dans des images. Dans le domaine 3D, nous avons résolu les problèmes d'appariement et alignement des vues partielles tridimensionnelles. Les résultats obtenus ont montré qu'avec notre approche, les performances sont nettement meilleures par rapport aux autres méthodes existantes.

Mots-clefs: description et appariement d'éléments caractéristiques, descripteurs locaux et globaux, sac-de-mots, mots visuels, recalage de vues partielles tridimensionnelles, information sémantique, information contextuelle.

ABSTRACT

This manuscript presents an extension of feature description and matching strategies by proposing an original approach to *learn* the semantic information of local features. This semantic is then exploited, in conjunction with the *bag-of-words* paradigm, to build a powerful feature descriptor.

The approach, ended up by combining local and context information into a single descriptor, is also a generalized method for improving the performance of the local features, in terms of distinctiveness and robustness under geometric image transformations and imaging conditions.

The performance of the proposed approach is evaluated on real world data sets as well as in both the 2D and 3D domains. The 2D domain application addresses the problem of image feature matching while in 3D domain, we resolve the issue of matching and alignment of multiple range images. The evaluation results showed our approach performs significantly better than expected results as well as in comparison with other methods.

Keywords: Feature description and matching, local and global descriptors, bag of words, visual words, range image registration, semantic information, shape context.

DESCRIPTEURS AUGMENTÉS BASÉS SUR L'INFORMATION SÉMANTIQUE CONTEXTUELLE

Par le biais de ce manuscrit, nous proposons une methode de description et d'appariement d'éléments caractéristiques des images. L'extension proposée a de meilleures performances en terme de caractère distinctif et d'invariance (robustesse) des descripteurs locaux dans les images.

L'approche suggérée est construite autour d'une technique originale, basée sur le paradigme de sac-de-mots pour apprendre la signification des éléments caractéristiques locaux qui conduit à une abstraction sémantique de la relation sous-jacente liée à plusieurs images.

Cette information sémantique est ensuite exploitée dans un sens contextuel, et en conjonction avec l'information locale pour genérer un descripteur d'éléments caractéristiques, qui est une combinaison linéaire de deux composantes, locale et sémantique contextuelle.

L'approche proposée, pour la description et l'appariement des éléments caractéristiques, se résume en quatre étapes:

1. *Description locale des éléments caractéristiques:*

   Des descripteurs locaux sont calculés pour un ensemble d'éléments caractéristiques (*e.g.*, point, régions, *etc.*) pré-sélectionnés, ou échantillonnés, sur des images. Le but est de capturer la distribution ou la variation de l'information au voisinage de chaque élément caractéristique choisi.

2. *Création de vocabulaires visuels:*

L'ensemble de descripteurs calculés précédemment, et collectés sur des images obtenues à partir de la même scène (ou sur des vues partielles représentant le même modèle 3D), sont correctement regroupés de manière à générer un nombre fixe de mots visuels. L' ensemble des centres de gravité des groupes résultants, représentent notre vocabulaires visuels

3. *Définition de contexte:*

   Chaque descripteur local est associé à un mot visuel, et une représentation de sac-de-mots est définie par le comptage du nombre d'éléments caractéristiques attribués à chaque mot. En particulier, pour un élément caractéristique de référence, son contexte est défini comme l'ensemble des sac-de-mots obtenus dans plusieurs régions qui définissent des coquilles concentriques centrées sur la référence.

4. *Description et appariement des éléments caractéristiques:*

   La mise en correspondance entre deux éléments caractéristiques est obtenue en comparant respectivement leurs composantes locales et contextuelles, et en prenant en compte les différents types de mesures.

Les résultats expérimentaux ont montré que la technique proposée est plus performante que les méthodes existantes. Pour les deux problèmes traités, dans les domaines 2D et 3D, les gains en performance les plus élevés sont enregistrés par nos descripteurs. Ceci est particulièrement illustré sur des images obtenues sous des conditions d'imagerie contraintes , où les descripteurs standard atteignent rapidement leurs limites et se révèlent inapplicables dans certains cas.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context and Motivation

Image description is an important task for many computer vision applications. These include, among others: object recognition [Ferrari 04, Lowe 04] and categorization [Opelt 04, Fergus 03], image retrieval [Mikolajczyk 01, Schmid 97], texture recognition [Lazebnik 05], feature matching [Tuytelaars 04, Schaffalitzky 02], making panoramas [Brown 03], video data mining [Sivic 03], robot localization [Se 02], 3D model registration [Huber 03, Makadia 06], and 3D feature matching [Mian 06].

The rationale behind image description is to provide a compact image representation, which is distinctive and invariant (robust) under geometric image transformations, imaging conditions, occlusion, and noises. To this end, different methods have been proposed, and these can be classified into three major categories, *local*, *global*, and *contextual* approaches:

- *Local:* for the local methods, an image is represented by a set of local descriptors, which encode the properties (*e.g.*, distribution, variation, etc) of the information collected in different feature neighborhoods.
- *Global:* based on global methods, an image is represented by a unique descriptor (vector) which encodes the properties (*e.g.*, distribution, variation, etc) of the information available on the whole image.
- *Contextual:* contextual methods span both local and global methods. For instance, shape context [Belongie 02] is a description method for which an image is represented by a set of local features (similar to local methods) by using the information collected on the whole image scope (similar to global methods).

Basically, a successful image description method has to perform indifferently while maintaining good performance (*i.e.*, distinctiveness and robustness) on different types of images and under challenging image alterations. These include, for example, the problems that arise with images extracted from homogeneous and textured scenes, or with those subjected to complicated geometric transformations.

To approach these problems, several techniques have been suggested. In the seminal work of Mikolajczyk [Mikolajczyk 05a], a number of promising local feature description approaches, including the contextual approach of Shape Context [Belongie 02], are evaluated and compared for image feature matching. The results suggest SIFT [Lowe 04], PCA-SIFT [Ke 04] and GLOH [Mikolajczyk 05a] as the most successful descriptors.

For object recognition, the evaluation conducted by Bay et al. [Bay 06] shows their SURF descriptor performs better than SIFT, PCA-SIFT and GLOH.

Recently, an interesting framework [Maji 09] has been elaborated in the context of object recognition to compare the performance of number of methods widely used for image description. These are evaluated on the popular *Caltech 101* dataset [Fei-Fei 06].

The results show that the performance of SIFT, Geometric-Blur [Berg 01], and Shape-Context descriptors is superior to others. Besides, GLOH works poorly in spite of its high performance on Mikolajczyk's dataset. It is also illustrated that the performance of SIFT is highly correlated with Shape-Context but the correlation dips with Geometric-Blur.

On the other side, the recent local shape (*i.e.*, for 3D model) descriptors of [Heider 12, Tang 12] based on distance to plane, normal distribution, mean curvature, Gaussian curvature, and shape index, are shown to be well adapted for the tasks related to 3D shape models like recognition and categorization of 3D objects.

Despite their attractive usefulness, it seems no approach of the aforementioned including others (or category of approaches) was found to perform best for all the image types, deformations, and tasks.

For instance, some problems with the approaches based on the local information, lack in performance for scenes exhibiting self similarities (*e.g.*, homogeneous-structured and highly-textured environments) as well as those depicting complicated non-affine distortions and non-rigid movements.[1]

Hence, it becomes difficult for matching features within images obtained from these scenes assuming affine warps or 2D-rigid transformations. By allowing non-affine image transformations, the local descriptors may fail to address, for example, the matching and registration problems. These constraints turn quite problematic for applications demanding accuracy and precision.

The alternative solution to overcome the limitation of the local descriptors in images with multiple similar regions, is probably adopting a contextual descriptor. However, this is far to be suitable for images presenting occluded points (like in 3D partial views) or altered by some geometric deformations, *e.g.*, scale change. This is because the contextual methods are more useful for tasks which the emphasis is more upon comparing shapes than on matching features.

The other solution involves using a global approach is suited only for the problems related to distinguish between shapes, but not for those of features discrimination. For example, this should be a good solution for tasks like object categorization and 3D shape retrieval, whereas it is not adapted for others like those involving estimating geometric transformations between images.

[1] These are often caused by non-stationarity of objects inside images, *i.e.*, objects move independently during image capturing or deformation.

To deal with all the aforementioned problems –among others, we propose an approach based on combining local and context descriptors.

Combining local and context information is a promising approach, and only few methods [Carneiro 04, Mortensen 05] adopt this strategy, to the best of our knowledge (see [Mian 05], for an extensive overview of 3D shape matching methods).

An interesting and effective approach has been proposed in [Frome 04], which is an extension of the so called *shape context* [Belongie 02] to the 3D domain.

Shape context encodes the distribution over relative positions of a fixed point with all the other points of the shape. In this fashion, it summarizes the global shape in a rich local descriptor [Belongie 02].

In this manuscript, we improve the basic idea of the shape context and thus we propose an original approach for combining local descriptors with the Bag-of Words (BoW) paradigm. This can be generalized to any local descriptor. Also, it can be seen as a generalization of the closest existing methods of [Carneiro 04, Mortensen 05].

In this work, we introduce an original approach which can be generalized to any local feature descriptor (*i.e.*, can be adopted with any local feature descriptor) to build powerful feature descriptors. The outlines of our strategy for describing and matching image features, are presented in the following section.

1.2 Semantic-Shape-Context Approach

The proposed Semantic-Shape-Context is a four-steps process, which involves both feature description and matching. This process will be detailed in Chapter 3, and here it can be outlined as follows:

1. *Local feature description:* Several local feature (*e.g.*, point) descriptors are computed in order to capture the local image (or shape) variation [Lowe 04, Johnson 99, Petitjean 02] in the point neighborhood.

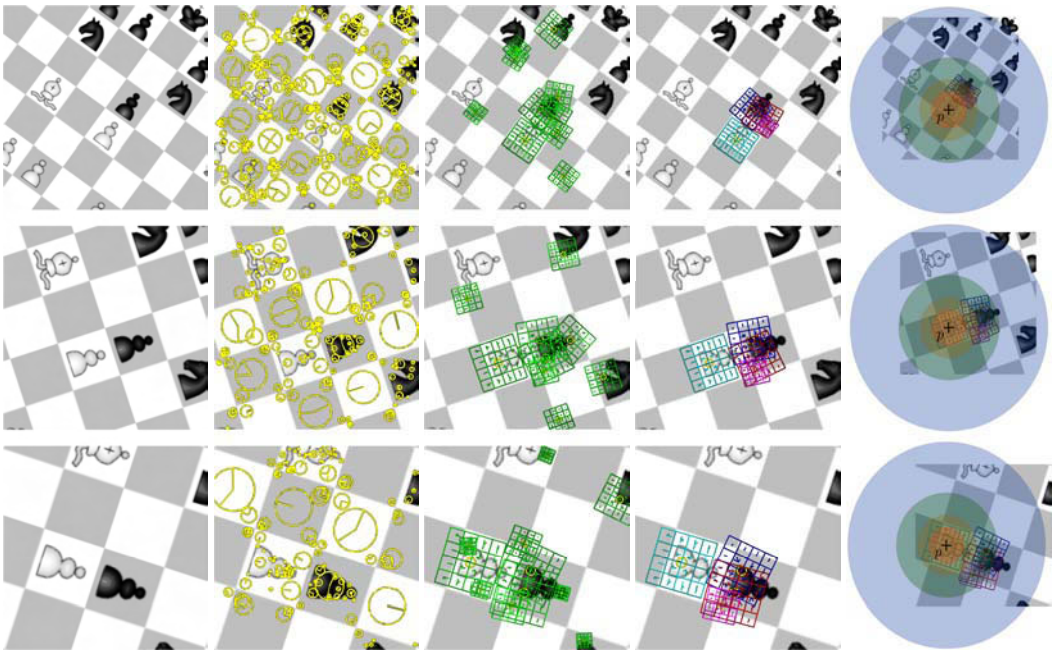
2. *Visual vocabulary construction:* The set of feature descriptors collected from all the images of the same scene (or from all the views of the same 3D model) are properly clustered in order to obtain a fixed number of visual *words*. Thus, the set of cluster centroids represents words in a visual dictionary [Csurka 04].

3. *Context definition:* Each local descriptor is assigned to a visual word, and a BoW representation is defined by counting the number of points assigned to each word. In particular, for a fixed point its context is defined as the set of BoWs computed

on several regions which are defined by concentric shells centered on the fixed point itself.

4. *Feature matching:* The matching between two features (*e.g.*, points) is computed by comparing their respective signatures and by taking into account the different kinds of descriptors. Both the *local* and *contextual* contributions are considered.

An illustration of the first three of the aforementioned steps is given in Fig. 1.1c, Fig. 1.1d , and Fig. 1.1e, respectively.



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

FIG. 1.1: Principle of semantic-shape-context feature description approach. Given a set of images (a), firstly, features are selected on each image, as shown in (b). Then, the proposed feature description approach is performed across three steps: (c) *local feature description*, (d) *visual vocabulary construction*, and (e) *context definition*. This illustration is based on features (*i.e.*, yellow circles) and descriptors (*i.e.*, green grids), which are computed using SIFT detector and descriptor.

The underlying idea here consists in the fact that the proposed context encodes not only the spatial relationship between features, but also their *class* with respect to each local descriptor. This means features assigned to the same cluster belong to the same class.

We thus call this new representation: *Semantic Shape Context* (SSC). The term *semantic* is used to emphasize the fact that we learn the local shape of the feature, where here the semantic is inferred by the feature classification.

It is worth noting that the choice of local point descriptors is not the focus of this work, since in principle any set of local descriptors can be used and cast in the proposed context. The effectiveness of the SSC is tested in both the 2D and 3D domains by addressing two problems of image feature matching and alignement of multiple range images[2] (*i.e.*, 3D multiview surface matching and registration).

1.3 Roadmap

The remainder of this manuscript is organized as follows:

In Chapter 2, we present a literature review of the most relevant approaches already published on the topic of image description, for both the 2D and 3D domains.

The main steps of the proposed feature description and matching strategy are presented in Chapter 3.

In Chapter 4, we illustrate the performance expected of Semantic-Shape-Context to be powerful concept for computing augmented feature descriptors.

The experimental results for evaluating the concept of Semantic-Shape-Context to resolve a 2D-domain problem (*i.e.*, image feature matching), are given in Chapter 5.

The proposed concept is also evaluated in 3D-domain by addressing the issue of matching and registration of multiple range images. This is reported across Chapter 6.

Finally, the conclusion is given in Chapter 7.

[2]This implements a fully automatic model registration pipeline matching framework

Chapter 2

Related Work

2.1 Introduction

Image description is often an early step, and sometimes the main step, for several machine vision tasks. These include, for example, video tracking [Trucco 06, Gabriel 03, Noldus 01], robotic mapping and navigation [Thrun 02, Thrun 00], object recognition [Lowe 99, Belongie 02], scene classification, and features matching [Ke 04].

Feature matching between two images is a typical case in which descriptors are computed on a number of pre-selected features.

Basically, the performance, in terms of precision, of any matching algorithm is directly proportional to feature accuracies, distinctiveness, and invariance of descriptors. Many feature description methods have been suggested to achieve the performance sought for high discriminative and invariant descriptors, and aiming to provide an accurate feature matching.

According to the number of descriptors (vectors) involved in the image description, the different image description approaches can be classified as *locals* and *globals* methods. The contextual methods presented early in the introduction (*i.e.*, Chapter 1) as a separate category, are included here with local methods.

Following is a literature review on the most relevant approaches already published on the topic of image feature description, in both the 2D and 3D domains, and with respect to the local and global categories. Besides, we will present in Section 2.4 a brief summary of a number state-of-the-art comparative studies designated for evaluating the performances of different feature descriptors.

Recaps of the most relevant 2D and 3D feature descriptors, among these reviewed here, are given in Tab. 2.1 and Tab. 2.2, respectively.

2.2 2D Feature Descriptors

Many 2D-domain vision problems are heavily dependent on the early task of image description. Video tracking [Trucco 06], robotic navigation [Thrun 02], object recognition [Belongie 02], and features matching [Ke 04] are few examples.

In this context, different approaches have been proposed. These can be partitioned, as above-mentioned, into two classes: local (including contextual) and global methods.

In general, the global approaches are useful for problems involving image shape comparison (*i.e.*, object recognition) whereas for those related to image feature matching, *i.e.*, estimating of geometric transformations, the local and contextual methods become

more appropriate.

2.2.1 Local Approaches

A basic contextual descriptor is a two-dimensional histogram representing the relative distribution (*i.e.*, relative to a reference point) of interest points inside an uniform square-grid.

Based on a close idea, instead of interest points, Shape Context [Belongie 02] is computed as log-polar histograms for spatial distribution of edges, extracted using Canny detector [Canny 86]. This technique has been successfully evaluated in shape recognition, in which edges are reliable features.

Spin image [Johnson 99, Johnson 97a, Johnson 97b] is an approach developed in the context of 3D object recognition. A 2D version (for 2D images) of this method has been introduced by Lazebnik et al. [Lazebnik 05]. The proposed *intensity-domain spin images* (abbreviated here as Spin-Image) descriptor is inspired by the latter 3D spin images, in which the traditional coordinates (*i.e.*, radial-distance and signed-elevation scalers are replaced by the point location and brightness respectively.

The spin image histogram is obtained as a function of 4-bins distance and 10-bins intensity, thus leading to a descriptor vector of dimension 40. The intensity-domain descriptor has a high degree of invariance for representing affine normalized patches. This is because the parameters, the distance from the center point and the intensity value are invariant to orthogonal transformations of an image. Fig. 2.1 illustrates the basic idea of building the intensity-domain descriptor based on modified spin images.

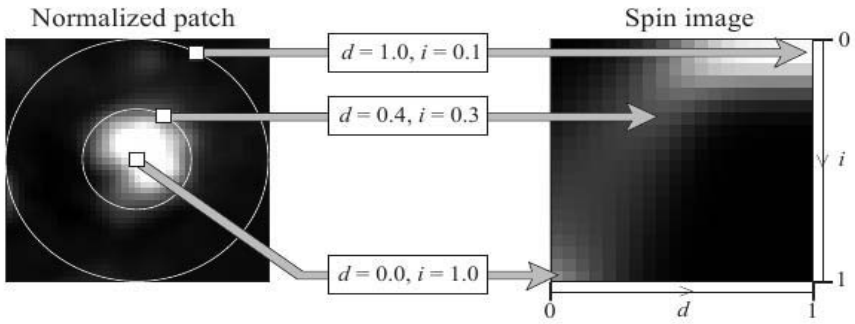


FIG. 2.1: Construction of the intensity-domain spin-image descriptor. The dimensions of each (*right*) descriptor histogram are: d and i, which are the distance from the feature-point and the brightness-value respectively. The slice of the spin image corresponding to a fixed d is simply the histogram of the intensity values of pixels located at a distance d from the center. This figure shows an example of three points on the normalized patch (*left*) and their different corresponding positions in the descriptor (*right*). (adapted from [Lazebnik 05]).

An approach robust to illumination changes [Zabih 94] was developed by exploiting the ordering and reciprocal relations between pixel intensities. Its histograms represent a distribution of all possible binary string combinations. These binary strings encode binary relations between intensities of several pixel in the neighborhood. This descriptor performs well in case of texture representation. However, it needs high dimensionality to build a reliable descriptor [Ojala 02].

Complex filters [Schaffalitzky 02] is a differential-based descriptor. It is developed in the context of multi-view matching (*i.e.*, establish relative view-points) given a large number of images where no ordering information is provided. First step of this approach is to normalize the intensity power in the neighborhood to unity, after shifting the signal mean to zero. This guarantees invariance to illumination changes. Based on the idea proposed by Baumberg [Baumberg 00], invariance of complex filters to image transformation is achieved by mapping the neighborhood onto a unit disk. The filters used are derived from the following model:

$$K_{mn}(x,y) = (x+iy)^m (x-iy)^n G(x,y), \tag{2.1}$$

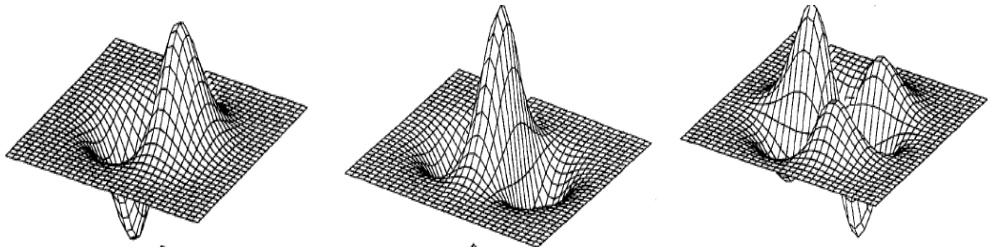
where $G(x,y)$ is a Gaussian. The original implementation used a total of 16 complex filter responses per image patch. This is obtained by setting: $m+n \leqslant 6$, $m \geqslant n$, and swapping m and n to obtain complex conjugate filters. The filter bank shown above differs from a bank of Gaussian derivatives in sense that linear coordinates change in the space of the filter response.

Differential invariants [Koenderink 87] and steerable filters [Freeman 91] are two descriptor close to complex filters, and use derivatives obtained by convolution with an approximated Gaussian. Differential invariants based on the property that the derivatives of the blurred illumination are equal to the convolution of the original image with certain filters of RF (*i.e.*, receptive field, and it is similar to Gabor filter [Gabor 46]). The framework termed this as *fuzzy derivatives*. The RF filters are derived from the following family of equations:

$$\varphi_n(x;t) = \frac{\partial^n}{\partial x^n} \frac{e^{\frac{-x^2}{4t}}}{\sqrt{4\pi t}} \tag{2.2}$$

To build the descriptor, a set of RF filters are concatenated to obtain higher order derivatives at lower resolution. This is done by exploiting the *concatenation theorem*. The filters use local jets [Poston 96] up to fourth order to compute edges curvatures. Examples of RF filters are shown in Fig. 2.2.

The steerable filters [Freeman 91] technique is similar to the differential invariants. It uses oriented (steered) filters to compute derivatives in an arbitrary direction. Oriented filters are a linear combination of basis filters. Responses of the basis filters are used to



(a) (b) (c)

FIG. 2.2: An example of RF filters. These filters play different roles: (a) This filter is used to detect edges or lines. In case for edges detection, the filter (b) plays the role of the curvature sensitive element for boundary curvature detection, whereas it is the curvature sensitive element for line curvature if (a) is used to detect lines. In this case (c), it is the curvature sensitive element for boundary curvature. (Illustration inspired by [Koenderink 87])

determine the direction of the derivatives. These derivatives are invariant to rotation if they are computed in the direction of gradient. Fig. 2.3 shows an example of steerable filters computed from circularly symmetric Gaussian and applied on a patch of a circular disk.

Affine and photometric moment invariant [Van Gool 96] is a method developed in the context of viewpoint invariant recognition of planar pattern. It uses traditional geometric moments as the basic features. The idea of this approach is based on moment invariants [Flusser 09]: *(i)* mixing of different types of moment and using a combination of coplanar patterns and *(ii)* to deal with invariance to illumination changes, the intensity moment is incorporated in the mixed moment. The moments are up to the second order (*i.e.*, keep the order of moments low), since high orders introduce more noise, that is, sensitive to small geometric and photometric deformation. Two kinds of moments are computed:

$$MS_{pq} = \iint_{\Omega} x^p y^q dxdy \quad \text{and} \quad MI_{pq} = \iint_{\Omega} I(x,y) x^p y^q dxdy \tag{2.3}$$

These are the *shape (p,q)-moment* and *intensity (p,q)-moment* of order $p+q$ respectively. The moment invariant is function of both shape and intensity moments. The descriptor based on moment invariants performs well with color images because the invariant can be computed for each color channel and between channels [Mikolajczyk 05a].

Cross-correlation is a basic descriptor represented by a vector of image pixels. This vector can be used in many tasks such such as image feature matching, pattern recognition, and feature detection [Gonzalez , Duda 98]. The traditional normalized correlation operation, used in cross-correlation, does not meet speed requirements for time-critical applications. Therefore, and due to the computational cost of spatial domain convolu-

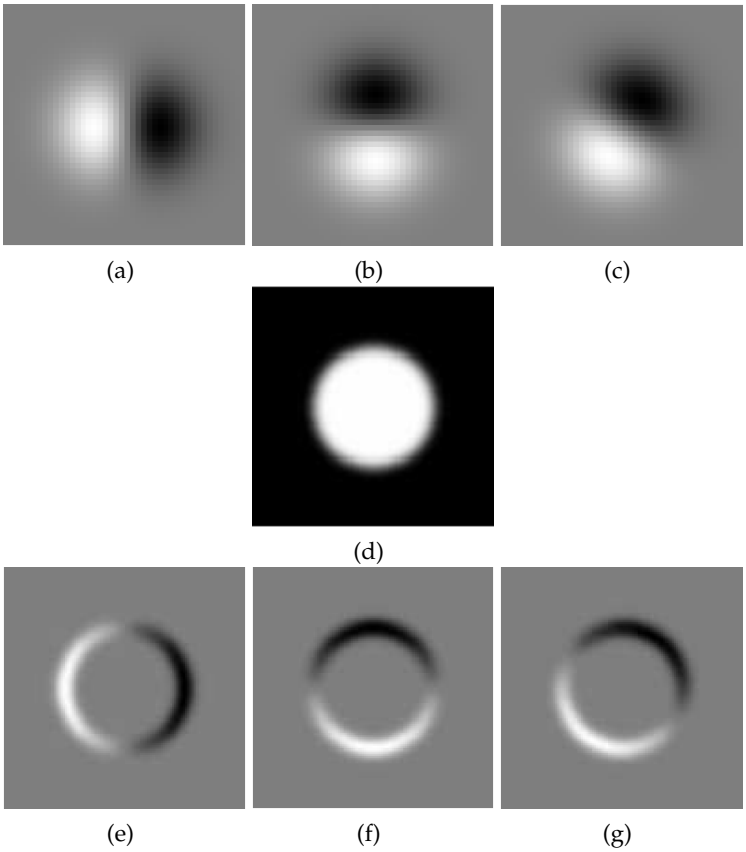


FIG. 2.3: An example of steerable filters with a circularly symmetric Gaussian kernel,G. (a) The first derivative along the positive x-axis (horizontal direction), $G_1^{0^\circ}$. (b) The first derivative along y-axis (vertical direction), $G_1^{90^\circ}$, that is, $G_1^{0^\circ}$ rotated (steered) by 90°. (c) The directional first derivative in the direction of the vector making an angle of 30° from the positive x-axis, $G_1^{30^\circ}$. This can also be computed as a linear combination of $G_1^{0^\circ}$ and $G_1^{90^\circ}$: $G_1^{30^\circ} = \frac{1}{2}G_1^{0^\circ} + \frac{\sqrt{3}}{2}G_1^{90^\circ}$. (d) A circular disk patch sampled from an image. The convolution results of this patch using $G_1^{0^\circ}$ and $G_1^{90^\circ}$ are shown in (e) and (f) respectively. (g) The patch convolution with $G_1^{30^\circ}$ can be computed directly as a linear combination of the convolved patches obtained in and (e) and (f), that is, $(g) :: patch = \frac{1}{2}(e) :: patch + \frac{\sqrt{3}}{2}(f) :: patch$. (Illustration inspired by [Freeman 91])

tion, the descriptor is not suitable for feature tracking task. One solution is computing correlation in the frequency domain using the FFT (Fast Fourier Transform). However, for template matching, the normalized form of correlation preferred in this task, has not a simple and efficient frequency domain expression. For this reason, many algorithms have been introduced to overcome this constraint. In the framework, *first template matching* [Lewis 95], the unnormalized cross-correlation is normalized using pre-computed tables containing the integral of the image and its square image over the search window. This descriptor is well suited for special effects feature tracking.

Briechle and Hanebeck [Briechle 01] proposed fast normalized cross correlation in the context of template matching. The algorithm uses a sum expansion of a given template function and a rectangular basis function. It is useful and suitable for problems where many different template are to be found in the same image.

A new approach has been proposed by Yoo and Han [Yoo 09] recently to compute the normalized cross-correlation without using multiplications. They showed that for a search window of size M and a template of size N the fast normalized cross-correlation requires only approximately $2 \times N \times (M - N + 1)$ additions or subtractions.

SIFT-like methods are distribution-based descriptors. Generally used in the context of 2D image feature description, where they have been proven to be very successful in many applications. These descriptors are derived from scale invariant feature transform (SIFT) [Lowe 04], which is a scale invariant region detector and descriptor based on the distribution of gradient magnitudes. The detector finds interest points at particular scales with assigned orientations. This keeps the detected points invariant to image location, rotation and scale. The descriptor is then computed as a set of orientation histograms (orientation relative to the interest point) on 4×4 pixel neighborhoods. Each descriptor contains an array of 4 histograms around the interest point, and each histogram contain 8 bins. These histograms are computed from magnitude and orientation values. The gradient magnitude and a Gaussian of scale σ (set to 1.5 times of interest-point scale) are used for weighting the contribution of each pixel in histograms. The resulting descriptor is a vector of dimension 128, *i.e.*, $4 \times 4 \times 8 = 128$ elements. This vector is then normalized to enhance invariance to changes in illumination. In order to reduce the effect of non-linear illumination , the vector is renormalized using a threshold of 0.2. An example illustrating the principle of SIFT approach is shown in Fig. 2.4.

Performance evaluation conducted by Mikolajczyk and Schmid [Mikolajczyk 05a] for different approaches has shown that SIFT is partially invariant to the image distortion (like viewpoint and illumination) and highly distinctive. They have been illustrated that in the context of feature matching the accuracy (measured by recall and precision) for viewpoint change of 50 degrees is higher than 50%. They also tested the distinctiveness through varying number of features in the database [Mikolajczyk 05a]. This

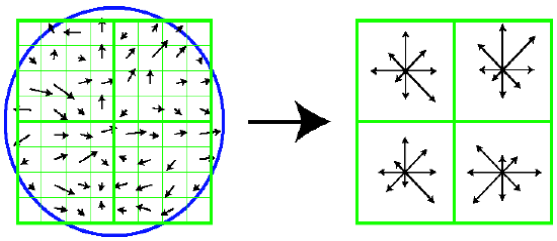


FIG. 2.4: An example of SIFT descriptor represented with a 2×2 array. The standard SIFT is 4×4 descriptor array computed from a 16×16 samples instead of 8×8 as shown in this example. The descriptor computes first the gradient and orientation at each image point neighborhood as shown on the left-side. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4×4 subregions, as displayed on the right-side, where the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region (adapted from [Lowe 04]).

demonstrates SIFT a very distinctive descriptor even with large number of features.

Gradient location and orientation histogram (GLOH) [Mikolajczyk 05a] is an extension of SIFT descriptor. It outperforms SIFT and other descriptors in sense that it increases the robustness and distinctiveness. It uses SIFT computed for log-polar grid location (changing location grid of SIFT) with 3 bins in radial direction and 8 bins in angular direction. The gradient orientations are divided in 16 bins. The results in an histogram of dimension 272 bins which is then reduced to 128 using PCA (*i.e.*, principal component analysis). The covariance matrix of PCA is computed from a set of 47,000 image patches.

PCA-SIFT [Ke 04] is another SIFT-like descriptor. It is a standard SIFT processed through the principal components analysis, in which the descriptor vector is computed within region of image gradients in x and y directions. The vector is of dimension 3,042, and reduced to 36 using PCA.

Colored SIFT (CSIFT) [Abdel-Hakim 06] is a version of SIFT descriptor which exploits color invariant characteristic in images. It is built around the photometric reflection model derived from the Kubelka-Munk theory [Geusebroek 00, Geusebroek 01]. This is based on the following model:

$$E(\lambda, \boldsymbol{x}) = e(\lambda, \boldsymbol{x})(1 - \rho_f(\boldsymbol{x}))^2 R_\infty(\lambda, \boldsymbol{x}) + e(\lambda, \boldsymbol{x})\rho_f(\boldsymbol{x}) \tag{2.4}$$

Here, the reflected spectrum in the viewing direction, E , if a function of the illumination spectrum, e, the Fresnel reflectance, ρ_f, and the material reflectivity, R_∞. The variables $\boldsymbol{x}$ and λ denote a point position in image and the wavelength respectively. To make the descriptor robust to different geometrical transformations and photometric changes, SIFT descriptor is computed with color gradients instead of gray gradients. The descriptor was evaluated in the context of feature matching for illumination changes

and using low resolution images with sizes of the order of 384×288. It showed significant improvements compared to the classical SIFT in terms of repeatability and number of correct matches.

Taboone et al. [Tabbone 06] presented a descriptor for object recognition. It is based on Chamfer distance [Borgefors 84] and the new $\mathcal{R}$-transform derived from the Radon transform [Deans 83]. The Radon transform of an image $I(x,y)$ is computed as follows:

$$T_{R^I}(\rho,\theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I(x,y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy, \tag{2.5}$$

where δ is the unit-impulse function. They introduced $\mathcal{R}$-transform of an image $I(x,y)$ as follows:

$$\mathcal{R}_I(\theta) = \int_{-\infty}^{+\infty} T^2_{R^I}(\rho,\theta)d\rho. \tag{2.6}$$

The $\mathcal{R}$-transform-based descriptor is tested in the context of shape recognition and it showed to be invariant to the translation and scale changes.

Tuzel et al. [Tuzel 06] have proposed using the covariance matrix of the interest region for computing the feature descriptor. They defined the region covariance of an n-dimensional feature as the norm of the first and the second intensity derivatives along x-axis and y-axis for all n components. They derived a fast method to calculate the covariances based on integral images in which the covariances are obtained by few arithmetic operations involving generalized eigenvalues. The algorithm performance was evaluated for object detection and texture classification. The results showed that the descriptor invariance resists to rotations and illumination changes.

In the context of action recognition in video sequence, Kläser et al.[Kläser 08] investigated an approach derived from HOG-based (Histogram of Oriented Gradient) representation [Dalal 06]. The representation is computed for the oriented 3D spatio-temporal-gradient (3D-gradients). This is similar to SIFT-like histogram [Lowe 04]. The approach presents a descriptor in which, videos are seen as spatio-temporal volumes, the HOG is extended to 3D domain, and integral image concept is extended to integral videos. An original 3D orientation quantization method based on regular polyhedrons is also proposed. The focus of the descriptor is action classification. Fig. 2.5 summarizes the procedure in building a spatio-temporal descriptor.

The latter descriptor was evaluated on different datasets, the KTH [Schuldt 04], Weizmann [Blank 05] and Hollywood [Laptev 08] actions and also compared to a number of descriptors, the Local Jets [Schuldt 04], Gradient+PCA [Wong 07], HoG [Laptev 08] and HOF [Laptev 08]. The spatio-temporal descriptor showed to be better in terms of action recognition accuracies.

DAISY [Tola 08] is a descriptor derived from SIFT and GLOH. Its shape is close to that introduced by Winder and Brown [Winder 07]. DAISY descriptor showed to be well

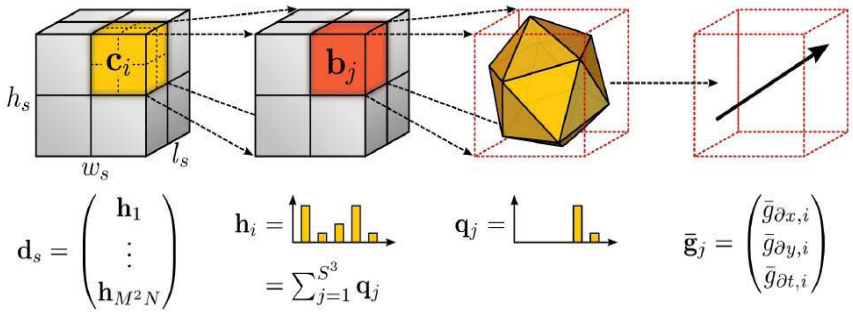


FIG. 2.5: The main steps for computing the spatial-temporal descriptor. (*1st column*) In the neighborhood of an interest point $\mathbf{C}_i$, a number of $2 \times 2 \times 2$ gradient orientation histograms are built and then concatenated to compute a full spatio-temporal descriptor $\mathbf{d}_s$. (*2nd column*) A grid of mean gradient of size $2 \times 2 \times 2$ sub-blocks is used to compute a gradient orientation histogram $\mathbf{h}_i$. (*3rd column*) A regular polyhedron grid is used to quantize the gradient orientation. (*4th column*) A mean gradient is computed over the entire videos. (Diagram adapted from [Kläser 08])

suited for sparse matching but not for efficiency. It is designed for dense wide-baseline matching in which the computation needs to be much faster without introducing artifacts that decrease matching performance. Instead of using weighted sums as with SIFT and GLOH, sums of convolutions is used. This leads to low computational time. The performance of the descriptor was compared to those obtained by SIFT, SURF , NCC (Normalized Cross-Correlation)[Lewis 95], and Pixel Difference. It appeared that DAISY is much faster than SIFT, and produced fewer artifacts than the other descriptors.

A recent variant of the SIFT have been proposed by Toews and Wells [Toews 09]. The ranked-ordered SIFT (SIFT-rank) investigates the image descriptions for affine image matching using the *ordinal description* method. This method computes the image measurement in terms of their ranks in a ordered array, instead of the their raw values. That is, each histogram bin of the descriptor is computed as its rank in a sorted array. The ordinal description of N scalar-valued image measurements, $\mathbf{x} = \{x_1, \ldots, x_N\}$, is given by:

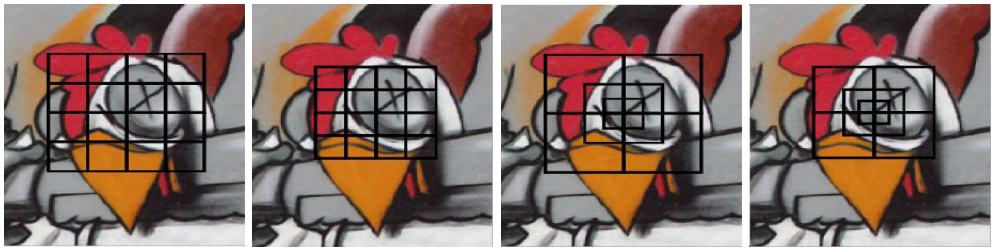
$$\mathbf{r} = \{r_1, \ldots, r_N\}. \tag{2.7}$$

Where the rank-order values r_i correspond to the values x_i are obtained as:

$$r_i = |\{x_k : x_k \leqslant x_i\}| \tag{2.8}$$

SIFT-rank descriptor is invariant against arbitrary monotonic variation in histogram bins. It is well suited for local-affine feature matching since it improves the performance of classic SIFT.

Irregular orientation histogram binning [Cui 09] is also a recent descriptor derived from standard SIFT. It uses an irregular histogram grid with subregions of different sizes. Unlike SIFT which encodes the gradient distribution into oriented histogram with a regular grid, histogram bin values are computed with respect to the feature center and over different subregion sizes. Fig. 2.6 shows examples of regular rand irregular subregions grid. The descriptor was tested in matching of local image features.



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

FIG. 2.6: Example of regular/irregular subregion grids. Regular sub-region grids used with SIFT, are shown in (a) and (b), while their corresponding irregular subregion grids adopted with *irregular-orientation-histogram-binning* descriptor are shown in (c) and (d). In case of (c) and (d), the image gradient is distributed over different subregions centered at the feature point. These subregions have different sizes and overlaps.

It showed a good robustness to scale quantization errors.

Fast and accurate descriptor has been introduced by Bay et al. [Bay 06, Bay 08]. Speedup robust features (SURF) is a couple of interest point detector and descriptor. The descriptor is computed using a previously selected orientation. This is determined based on the intensity distribution in the circular neighborhood of each interest point. Next, a square region is built in the direction of the selected orientation. The SURF descriptor is then 4-steps process obtained around this oriented square region. The process can be described in 4 stages:

1. Each pre-selected oriented square region of size 20s (s is predefined scale obtained from SURF detector), is divided regularly into 4×4 sub-square regions. Examples of pre-selected square regions are shown in Fig. 2.7a.

2. Haar wavelet is computed for each sub-square over a regular subregion grids of size 5×5 pixels. Splitting the regions into sub-regions using a regular grids and then computing the wavelet response are illustrated in Fig. 2.7b.

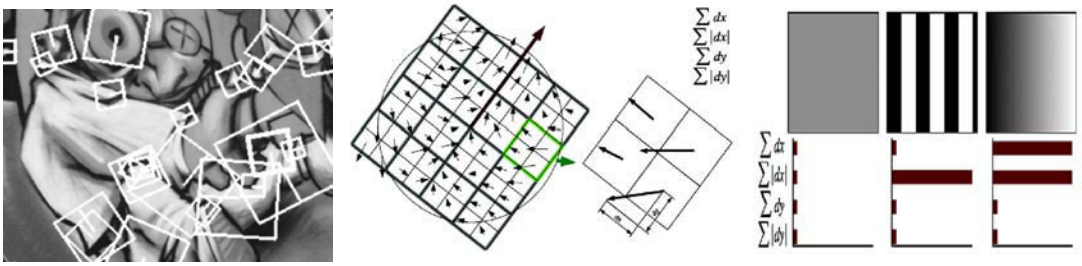
3. Both horizontal and the vertical wavelet responses, d_x and d_y, are weighted with Gaussian of scale, $\sigma = 3.3s$, and centered at the interest point. This augments the robustness to geometric distortion and localization errors.

4. The first two components of the descriptor vector are computed as the ingrates of the wavelet responses d_x and d_y over the sub-square regions. The two remaining components are set to sums of the absolute values, $|d_x|$ and $|d_y|$.

Thus, we obtain for each sub-square region a sub-vector of 4 components, $\sum d_x$, $\sum d_y$, $\sum |d_x|$, and $\sum |d_y|$. The resulting 64-dimensional SURF descriptor is obtained by concatenating all sub-vectors computed for 4×4 sub-square regions. The illustration of Fig. 2.7c demonstrates three different image patches with their corresponding SURF descriptors.

The SURF descriptor has good invariance against illumination changes, since the wavelet response is invariant to the variations in the illumination. In order to achieve the invariance to the scale change, each descriptor vector is normalized to unity.

Conjointly to SURF, an upright version (U-SURF) was proposed. U-SURF descriptor is speedy version well suited in case the invariance to the rotation can be neglected. It also increases the discriminative power.



(a) (b) (c)

FIG. 2.7: The principle of SURF descriptor. (a) Example of oriented square regions selected with different scales. (b) The wavelet responses are computed for each sub-square. SURF uses a size of 5×5 pixels for each sub-square, but here only 2×2 are displayed. The components of the descriptor are computed as the sums of wavelet responses, d_x, d_y, and their absolute values , $|d_x|$, $|d_y|$. The sums are over each sub-square region. This gives a descriptor of size 64 since there are 4×4 sub-square regions, and for each sub-square of size 5×5 pixels, 4 components are computed, $\sum d_x$, $\sum d_y$, $\sum |d_x|$, and $\sum |d_y|$. (c) Example of SURF descriptors computed for three different image patches.

CCH (Contrast Context Histogram) [Huang 06, Huang 08] is a global descriptor for image matching. The CCH encodes the contrast distribution in the interest point neighborhood into a 3D histogram using log-polar grid. This is similar to Shape-Context approach. The distinctiveness of the descriptor is enhanced through using both positive and negative contrasts. These are accumulated in separated bins in the histogram as shown in Fig. 2.8.

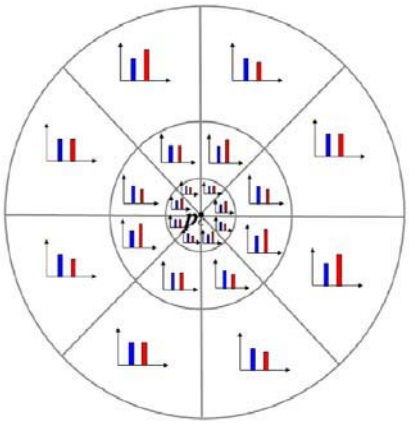


FIG. 2.8: Diagram of CCH histogram representation. For each sub-region inside a log-polar grid mapped in an interest point neighborhood, an histogram is constructed by computing and encoding the distributions of positive and negative contrasts in two separated bins. Here, the diagram shows blue and red bars as examples for positive and negative contrast bins.
(Image taken from [Huang 08])

The performances of the CCH descriptor was compared to those obtained with SIFT. The comparison is performed according to the computational time and matching accuracies. The results showed that CCH outperforms SIFT in the computational time complexity but approximately similar to it in matching accuracies.

2.2.2 Global Approaches

In general, the global descriptors are built around spatial-frequency techniques, which use histograms to encode different characteristics of appearance and shape[Ke 04, Belongie 02]. For example, the relative distribution of interest points around a reference interest point, which is computed inside a number of concentric shells and then represented by a 2D histogram. Based on a close idea, Shape-Context [Belongie 02] and geometric histogram [Ashbrook 95], instead of using points, exploit spatial distribution of edges to compute histograms. These techniques were successfully tested in shape recognition in which edges are reliable features.

Shape-Context is 3D log-polar histograms representing the distributions of edges extracted with Canny [Canny 86] detector. These distributions are computed with respect to log-polar coordinates, radial distance and orientation. The radial coordinate is quantized into 5 bins while the orientation into 12 bins. This gives a descriptor vector of dimension 60. Invariance to location is an intrinsic property of shape context because the distributions are computed with respect to a set of points on the image. Since all the radial distances are normalized by the mean distance, the invariance against scale

change is improved. A simple example for computing the shape context descriptor is illustrated in Fig. 2.9.

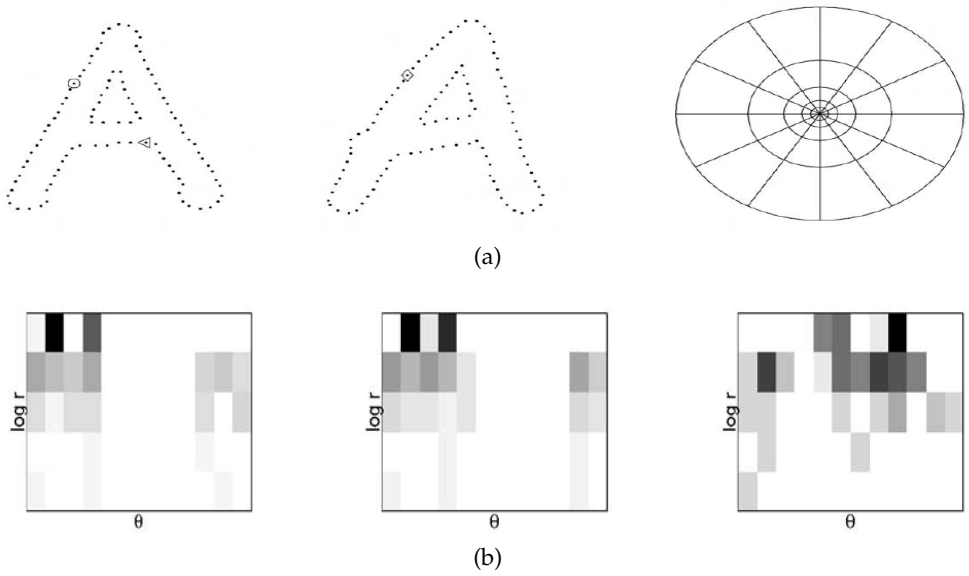


FIG. 2.9: A simple example for computing the shape context descriptor. (a)(*left* and *center*) Some edge points of two shapes and (*right*) the log-polar grid in which 5 bins of $\log r$ and 12 bins of θ are used. (b) Examples for shape contexts computed for points labeled $\circ$, $\triangleleft$ and $\diamond$ in the left and in the center of (a). Diagrams in (b) represent 3D histograms computed by accumulating edges points inside the log-plor grid shown in (a)(*right*). The accumulation is with respect to reference points such as $\circ$, $\triangleleft$ and $\diamond$ shown in (a). We should note here that a dark bin corresponds to strong accumulation of points, that is a large value in histograms. The shape contexts of points marked by $\circ$ and $\diamond$ appear to be similar. This because they are computed from similar points on the two shapes (adapted from [Belongie 02]).

Berg [Berg 01] suggested to adopt the geometric blur. The geometric blur denoted, $G_I(x,y)$, for an image $I(x,y)$ is defined as follows:

$$G_I(x,y) = \int_{W \in \Omega} I(W(x,y))dW \tag{2.9}$$

The integral is computed over a set of bounded warps Ω (geometric transforms). Fig. 2.10 illustrates the idea underlying the geometric blur technique. This consists in blurring the image by a spatially varying Gaussian filter to ensures that the blur should be smaller near the features points, and larger away from them. This is not the case with an uniform Gaussian filter.

The advantage of using the geometric blur rather than a simple uniform Gaussian blur is shown in Fig. 2.11.

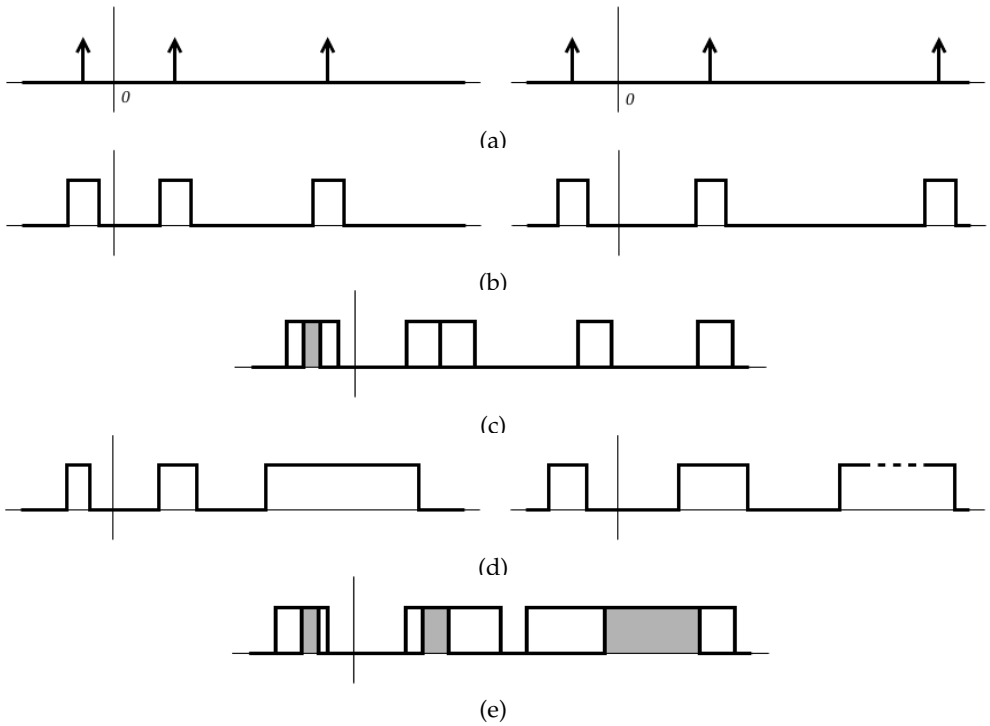


FIG. 2.10: An example illustrating the basic idea of the geometric blur approach. (a)(*left*) Three unit-impulse functions representing a 1D signal, $I(x)$. (a)(*right*) A translated version, $I_w(x)$, of $I(x)$. One wants to match positively between these two signals, whereas, for majority of translation amounts, the matching is obtained negative. As example the correlation gives often a null value. (b) One solution is blurring the signals with an uniform Gaussian filter. For a clarity purpose, here, Gaussian filters are shown as rectangle windows, and the window width corresponds to the filter size. (c) The resulting correlation using a Gaussian filter where the correlation between $I_w(x)$ and $I(x)$ decreases gradually as translation increases. This leads to an over blurring near the center, and under blurring away. (d) To overcome the latter problem, the geometric blur is used in which the amount of blur is proportional to distance from the origin (blur should be small near the corresponding points, and larger away from them). (e) Now after using geometric blur the correlation varies linearly.

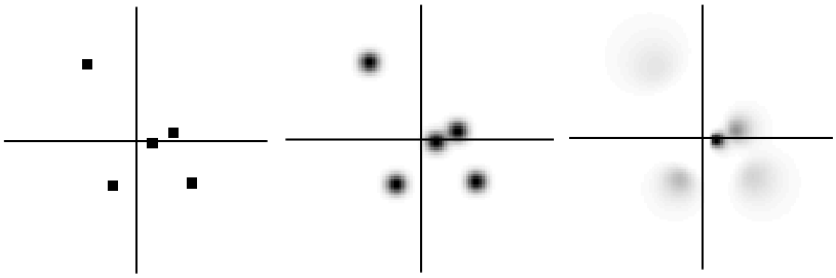


FIG. 2.11: An example to demonstrate the advantage of using geometric blur instead of approaches based on an uniform Gaussian blur. (a) In this coordinate system, a signal consists of a corresponding point chosen at the origin, and a set of five feature points in its neighborhood. (b) A blurred version of the signal shown in (a) obtained using an uniform Gaussian filter. This shows that all feature points are blurred with the same amount. (c) This version is obtained after applying the geometric blur on the signal shown in (a), for which the amount of blur is proportional to distance from the origin that is small near the corresponding point (the origin), and larger away from them (adapted from [Berg 01]).

To the best of our knowledge, only few methods exploit the idea of combining of local and context information. Carneiro and Jepson [Carneiro 04] have initiated this idea by combining local descriptors of [Carneiro 03] and SIFT with shape context. The proposed descriptor has been tested in the context of two applications, wide baseline stereo matching and non-rigid motion. The experiment results showed that the novel approach provides a higher inlier ratio than Hough clustering.

Later on, Mortensen et al. [Mortensen 05] suggested also to combine local SIFT with shape context to resolve ambiguities occurring in images containing multiple similar motifs. The descriptor is obtained as a weighted concatenation of local SIFT descriptor and shape contextual component:

$$\mathrm{D} = [w\mathrm{L} \;\; (1-w)\mathrm{G}], \tag{2.10}$$

where the parameter w is a relative weighting coefficient fixed to 0.5. The descriptor is of length 188, that is, 128 of local SIFT, L, and 60 of global context, G. The authors claimed that the descriptor is robust to local appearance ambiguity and non-rigid transformations. However, no convincing evaluation results are produced since no comparison to others approaches are provided.

Moreover, the evaluation was conducted only for one unique image, which is warped artificially for limited geometric transformations of rotation and skew only. The only experimental evaluations that the authors were able to provide are reported in Fig. 2.12. The few results they obtained showed that the combined-SIFT-Shape-Context performs better than SIFT and shape context taken separately. It appears clearly, without any doubt, that these results are not convincing and the approach is still far from being completely investigated.

Though the latter two approaches and our SSC seem to be close, they differ mainly in the way the context information (component) is *generated* and then *exploited*:

- Even though their meanings are similar they are not generated in similar manner. Whereas the other methods adopt the standard shape-context to encode the context information, we propose an original approach (*i.e.*, learning of the local shape of the feature) to generate the context information which encodes not only the spatial relationship between features, but also their class as a way for abstracting the semantic connection between images, and in principle, any local descriptor can be used and cast in the proposed context.
- In contrast with the other method, our strategy to match two image features consists in comparing first their respective local and contextual components. The distance between the feature descriptors is then computed as a weighted sum of both the distances between their respective local and contextual components. This

is more useful, since the metric (*e.g.*, euclidean distance and histogram comparison χ^2) to compute the distances between feature can be chosen [1] differently between local and contextual components. This also helps to reduce the dimensionality of each component, which in turn aids the matching to be more accurate and effective.

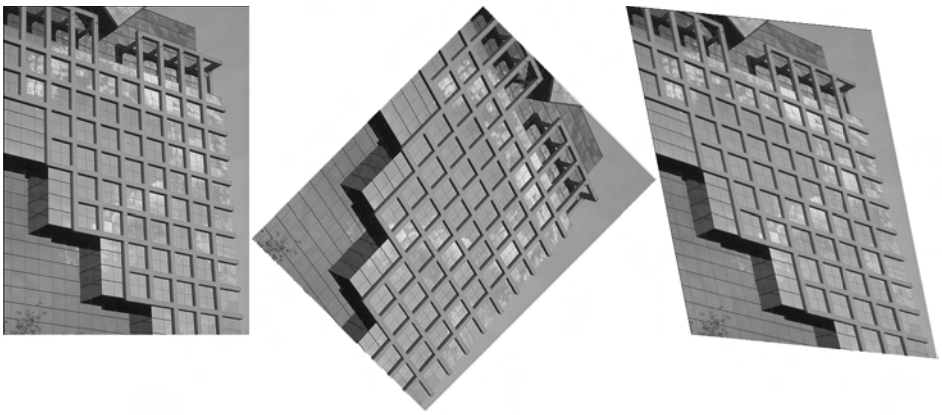
In addition to the previous arguments, the other methods use local and global components, which are completely decorrelated, without any notion of connection between images (the same visual word can appear on different images) like we propose. That is to say, they consider the context component, taken independent of the local component. Thus, the descriptors can deal with ambiguities resulting from the presence of multiple similar motifs in images, but not with other errors like those related to detectors, outliers, and occlusions.

Also, including all the above-mentioned approaches, much more emphasis has been placed on 2D image feature description and many other approaches have been introduced despite the methods cited above still gaining more interests.

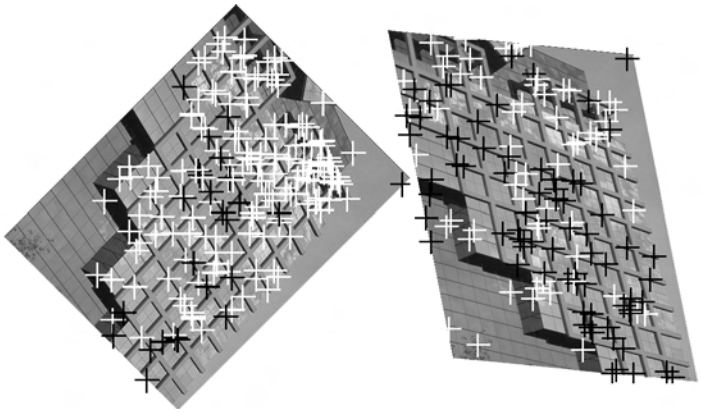
2.2.3 Recap

To summarize, the most relevant 2D feature descriptors (among the aforementioned) are recapitulated in Tab. 2.1.

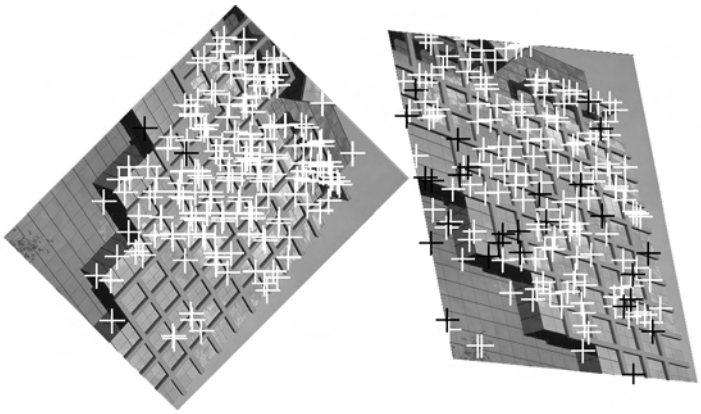
[1] For instance, adopting a metric of histogram comparison, like χ^2, for comparing the descriptors built around an histogram (or distribution) representation, will provide more accurate results than using, for example, the simple euclidean distance.



(a)



(b)



(c)

FIG. 2.12: Experimental results provided by the approach of Mortensen et al. [Mortensen 05]. (a) The image and its transformed versions used to evaluate the approach. Based on the nearest neighbor strategy, the authors reported the following matching scores: (b) SIFT only —rotation: 170/200 correct (85%), —skew: 73/200 correct (37%). (c) SIFT with global context —rotation: 198/200 correct (99%), —skew: 165/200 correct (83%). (Images and scores extracted from [Mortensen 05]).

TAB. 2.1: Recap of the most relevant 2D feature descriptors among the above-reviewed approaches.

Category	Name & Reference	Observation
Local	Spin image [Johnson 99]	Invariant to orthogonal image transformations
	Approach of [Zabih 94]	Performs well in case of texture images
	Complex filters [Schaffalitzky 02]	Invariant to illumination changes
	Differential invariants [Koenderink 87]	Provides different views point at the same time, from inside and outside
	Steerable filters [Freeman 91]	useful in analyzing image sequences and volumetric data
	Affine/photometric moment invariant [Van Gool 96]	Performs well with color images
	Cross-correlation [Lewis 95, Briechle 01, Yoo 09]	Well suited for special effects feature tracking
	SIFT [Lowe 04]	Partially invariant to the image distortion and highly distinctive
	GLOH [Mikolajczyk 05a]	Increases the robustness and distinctiveness of SIFT
	PCA-SIFT [Ke 04]	Best performance for image blur
	Colored SIFT (CSIFT) [Abdel-Hakim 06]	Improves the distinctiveness of SIFT under illumination changes
	Approach of Taboone et al. [Tabbone 06]	Invariant to the translation and scale changes
	Approach of Tuzel et al. [Tuzel 06]	Resists to rotations and illumination changes
	Kläser et al.[Kläser 08]	The focus of the descriptor is video action classification
	DAISY [Tola 08]	Well suited for sparse matching but not for efficiency
	SIFT-rank [Toews 09]	Well suited for local-affine feature matching
	Irregular orientation histogram binning [Cui 09]	Shows a good robustness to scale quantization errors
	SURF [Bay 06, Bay 08]	Has good invariance against illumination and scale changes
	CCH (Contrast Context Histogram) [Huang 06, Huang 08]	Outperforms SIFT in the computational time complexity
Global	Shape-Context [Belongie 02]	Invariance to location is an intrinsic property
	Geometric blur [Berg 01, Vyas]	Suited for template matching
Mixed local-global	Approach of Mortensen et al. [Mortensen 05]	Improves SIFT for rotated/sheared images containing multiple similar motifs

2.3 3D Feature Descriptors

In the context of 3D image registration, estimating rigid transformations that align corresponding points of range images (*i.e.*, 3D partial views)[2] is a critical issue for various tasks in computer vision, *e.g.*, 3D model reconstruction.

The ICP algorithm [Besl 92] is the gold standard for pairwise range image alignment. However, it requires a sufficient overlap among the range images and a coarse pre-registration to avoid getting stuck in a local minimum.

In particular, according to the taxonomy proposed in [Huber 03], when an initial estimate is unknown and more than two range images are involved, the problem is called *matching of multiple range images*. It consists of three main sub-problems need to be solved [Huber 03]. These include `(i)` finding out which range images are in overlapping, `(ii)` determining the relative pose between each pair of overlapping, and `(iii)` identifying the absolute pose of the range images.

Focusing on `(i)` and `(ii)`, similar to above for 2D feature descriptors, both local and global techniques are exploited.

In this context, the local techniques are based on point-to-point matching, in which each point signature describes local surface properties while the global techniques [Makadia 06, Vranic 01b] directly estimate the matching of the whole range images by comparing global surface characteristics.

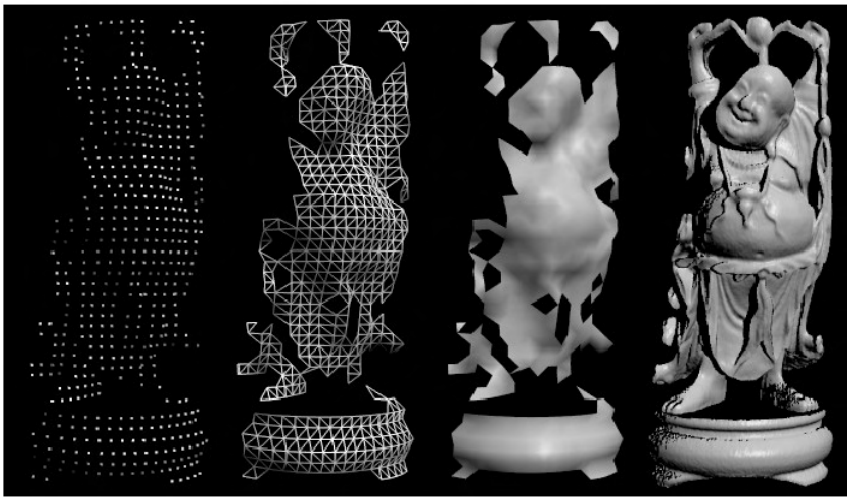
For a considerable number of 3D tasks, adopting either local or global approach, is often quite enough to achieve desired performances. However in particular tasks like those involving range images, the above focused problems turn to be challenging.

For instance, one of the challenging issues of the local approaches is the presence of multiple discontinuities, or *holes*, on the surface. This is a typical and common defect for many scanned 3D models. These defects are, in general, due to noise and systematic errors arising during creating range surface, as illustrated in Fig. 2.13.

Moreover, much as with the images in the 2D domain, the local 3D approach suffers from lack in performance for range images containing many regions with similar surface structures.

Though it seems in general to deal better against various problems, the global and contextual methods perform worst when the 3D shape contains a considerable number of points which diverge from the overall pattern. This is a frequent issue with range images for which the number of occluded points is usually high.

[2] In this thesis, the names *"range image"* and *"3D partial view"* are used interchangeably.



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

FIG. 2.13: An illustration of different steps to create a range surface. *(a)* A range image obtained by sub-sampling from 3D scanning-based device. *(b)* Connecting nearest-neighbor with triangular facets. *(c)* Performing shaded rendering. *(d)* Enhancing of the rang surface resolution. (Image taken from [Curless 99]).

To approach these problems, many techniques have been introduced and following is a literature review of the most relevant of them. Since the approaches reviewed here are mostly locals or contextuals, all the methods (including the globals) will be presented linearly, *i.e.*, within one section.

3D Spin Images [Johnson 97a] is considered so far, as a reference work for many other techniques. 3D Spin Images is a global approach based on the concept of object-oriented coordinate system.

The advantage of 3D Spin Image is that the oriented point is well defined at every point on the surface. Thus, it can be determined robustly, *i.e.*, unambiguously, almost at each point except at surface points discontinuities where the surface normal cannot be calculated. Precisely, at the points in which the first order surface derivative are undefined [Johnson 98].

Vranic et al. [Vranic 01a] proposed a method, built around 3D objects spatial properties, for describing 3D shapes. The idea is that, similar objects are represented by *close* points in the feature vector space. The descriptor is obtained as the coefficient absolute values of 3D discrete Fourier Transform, computed for coarse voxelization of a 3D model.

This approach is evaluated according to the discriminative power criterion in the context of 3D-model retrieval task, where its performances are compared to those

of three state-of-the-art methods. The experimental results showed the approach outperforms the others for rotation only. The authors claimed that the descriptor is also invariant to translation, scale, reflection, and is robust to level-of-detail. However, no evaluations are provided.

The 3D Shape Spectrum Descriptor (3D-SSD) [Zaharia 01] is an approach conceived for MPEG-7 Committee Draft (CD), and leads to produce an intrinsic shape descriptor of 3D meshes. It is based on the distribution of the local geometric information, shape-index, on the whole mesh.

The 3D-SSD is tested on MPEG6-7 3D model database of approximately 1300 meshes. The evaluation results of the descriptor in objective retrieval using ground truth of 15 categories containing 228 meshes, gives a percentage of Bull-Eye score of 85%.

In the context of content-based 3D model retrieval, Chen et al. [Chen 03] suggested a visual similarity-based 3D model retrieval approach (also called LightField descriptor (LFD)) to measure the resemblance between 3D models by using visual similarity. The principle of the proposed approach is that the similar models still appear like similar under different viewpoints. This is illustrated in the example of Fig. 2.14.

The method is evaluated and compared to three competing approaches, and it showed better scores, in terms of precision and recall criteria, of 42%, 94%, and 25%.

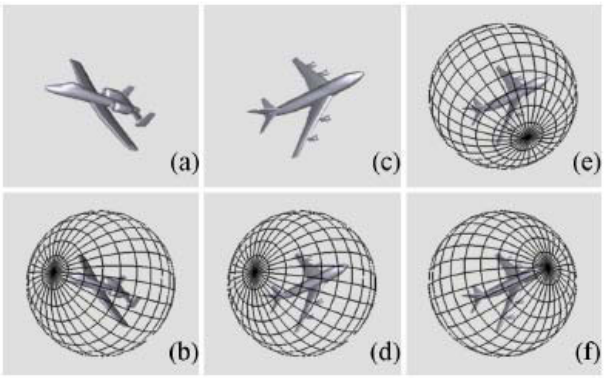


FIG. 2.14: An illustration of the idea of visual similarity-based 3D model descriptor. For different view angles, the two models of the first and second rows, remain look similar even though the angle of view changes. (Image taken from [Chen 03])

An approach invariant to rotation, based on spherical harmonics, has been proposed by Kazhdan et al. [Kazhdan 03] for matching and alignment of 3D models. In fact, the proposed descriptor is inspired by the approach of "*Fourier descriptors for plane closed curve*" [Zahn 72]. In other words, it is a generalization of the latter to spherical

functions. Fig. 2.15 illustrates the different steps in computing a rotation invariant descriptor of a spherical function.

This method is evaluated for classifying 3D models within a database of 1890 "household" objects. The performance comparison to other approaches, showed the proposed descriptor outperforming other methods while obtaining high matching performances, *i.e.*, *precision* and *recall* scores. Besides, the descriptor has a reduced dimensionality which makes it more efficient.

An interesting and effective approach has been proposed in [Frome 04]. It is 3D extension of 2D Shape-Context [Belongie 02] approach described in Section 2.2. The 3D Shape-Context, represented by a 3D histogram, which encodes the spatial distribution of points with respect to a reference point. It summarizes a global descriptors in rich 3D histograms. Thus, it is quite robust against outliers and shape defects.

In addition to the above-mentioned approach [Vranic 01a], Vranic [Vranic 05] suggested another 3D-shape descriptor called DESIRE. It is a composite of DEpth buffer images, SIlhouettes, and Ray-Extents of a polygonal mesh. Mathematically speaking, the composite feature descriptor is a concatenation of three feature descriptor components,

$$C = \begin{pmatrix} D & | & S & | & R \end{pmatrix} \tag{2.11}$$

To achieve an affine-transformation invariance descriptor, the triangle mesh models are transformed into canonical coordinate frames. This is obtained throughout: `(1)` shifting the center of gravity to the origin, `(2)` applying a rotation by using Continuous Principal Component Analysis (CPCA), `(3)` distance normalization (or scaling) to obtain the average distance (of vertices) to the origin equal to 1, and finally, `(4)` flip the vertices by using the moment test.

The proposed method is evaluated and compared to the aforementioned LFD approach [Chen 03], which is stated by the author as the best state-of-the-art descriptor. The experimental evaluations are performed on the BSP [Shilane 04] models of 1814 meshes belonging into 161 categories.

The results showed that the composite descriptor performs better than LFD in terms of retrieval effectiveness and computational time as well.

Mian et al. [Mian 06] proposed a novel approach for feature matching and automatic pairwise registration of range images. It is based on a new tensor representation which exploits third order tensors to represent semi-local 3D surface patches of range images. The authors claimed that the method is accurate and efficient. They also stated that, in comparison to 3D Spin-Images, it is more discriminative and performs better at low resolutions of range images.

Inspired by the standard SIFT approach for 2D image, the 3D-SIFT is a technique intro-

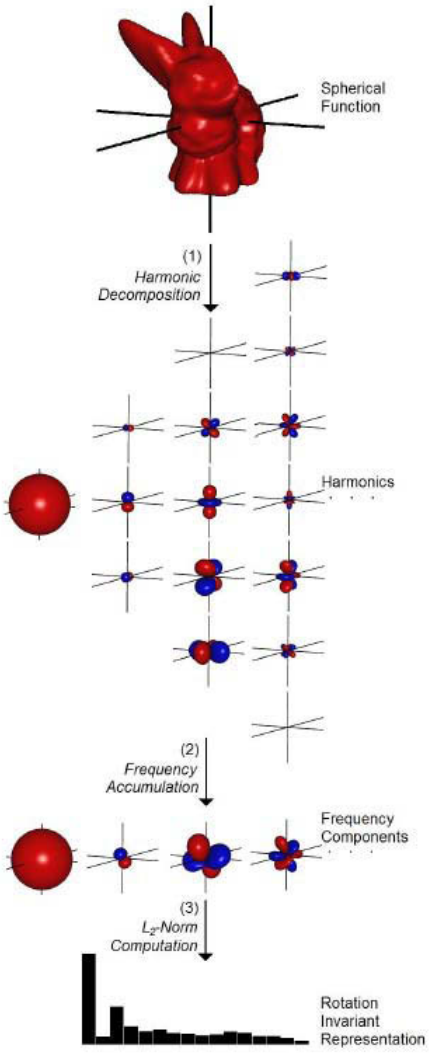


FIG. 2.15: Computation of a rotation invariant descriptor of a spherical function. This is three-step process: *(1)* the function is composed in its harmonics, *(2)* these harmonics are then summed for each frequency, and finally the norm of each frequency is computed. (Figure adapted from [Kazhdan 03])

duced by Scovanner et al. [Scovanner 07] in the context of video action recognition and 3D imagery. It uses Bag-of-Words paradigm to describe spatio-temporal relationships between video sequences.

The technique have been evaluated on the data set of Blank et al. [Blank 05], and dedicated for classifying actions within videos. The reported results showed the proposed descriptor performs relatively better than other state-of-the-art methods.

Another approach constructed around geometric scale-space analysis of 3D models, has been proposed in [Novatnack 08]. It consists of analyzing the scale-variability of range images and detecting the 3D features of scale-dependent. The resulting local 3D shape descriptors encode the local shape information within the inherent support region of each feature. The authors demonstrated that the proposed descriptors can be used in an efficient hierarchical registration algorithm for aligning range images with the same global scale.

A compact multiview descriptor is introduced by Daras and Axenopoulos [Daras 09] for 3D object retrieval. The method consists in generating a set of 2D images (multiviews) from a 3D object, and then a collection of 2D rotation-invariant descriptor is computed for each image. The authors asserted that the approach outperforms the other similar view-based descriptors.

Unique Shape Context (USC) is a descriptor proposed by Tombari et al. [Tombari 10] to improve the accuracy of 3D feature matching. The performance of USC is compared to that of the original 3D Shape Context, and showed this latter to be clearly outperformed by USC.

Recently, Tombari et al. [Tombari 11] presented also a combined texture shape descriptor for improving 3D feature matching. The suggested descriptor, named CSHOT, is a mixed histogram of normal orientations and texture-based information. This is illustrated in Fig. 2.16.

The CSHOT is evaluated for 3D object recognition in the presence of clutter and occlusions. The obtained performances are compared to those of SHOST [Tombari 10](*i.e.*, the original version of CSHOT) and MeshHoG [Zaharescu 09]. The results showed CHOST to be more accurate than SHOST and improves the efficiency of MeshHoG.

The CORS [Van Nguyen 11] – acronym for Concentric Ringing Signature – is a descriptor developed for 3D objects. The descriptor is computed based on the local geometric properties, which are preserved under continuous deformations of objects, *i.e.*, local topologies. The approach is illustrated in Fig. 2.17.

It is a three-steps method. First, a spherical support region is selected around a reference point p. Then, the local neighborhood of p is obtained as a plane, in which the normal

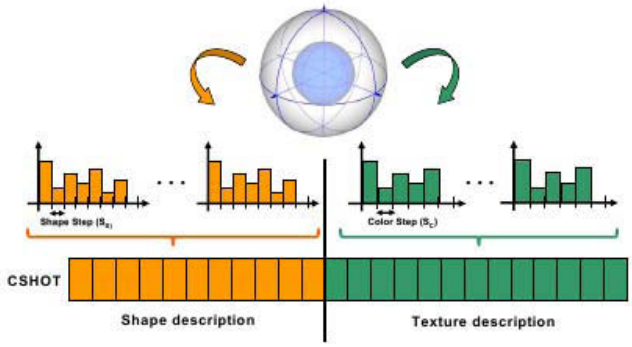


FIG. 2.16: A combined texture-shape descriptor is obtained by concatenating the shape-based and texture-based histograms (or signatures). (Image taken from [Tombari 11])

direction at the reference is adopted to be $z-axis$. Finally, the reference orientation $x-axis$ is selected and the distance from the surface to the patches are projected.

The experimental evaluations are performed on TOSCA and Mian data sets, based on the percentages of correct matches and recognition rates. Compared to 3D Spin Images and points signatures [Chua 97] approaches, CORS produces better results for both matching and 3D object retrieval. For instance, it increases the percentage of correct matches compared to the other methods, from 39% to 88%.

Nowadays, Kokkinos et al. [Kokkinos 12] introduce "*Intrinsic Shape Context*" (ISC) descriptor for deformable shapes. ISC is a meta descriptor usable with any shape property (photometric or geometric).

For 3D shape retrieval, the evaluation conducted by Tang and Godil [Tang 12] to compare a set of local shape descriptors inside the bag-of-words algorithm, suggests mean curvature, shape index, and curvature index, as the best descriptors.

Even though the above-mentioned techniques almost showed to work well under normal situations, their performances are rarely tested inside complicated environments like those described early in this section.

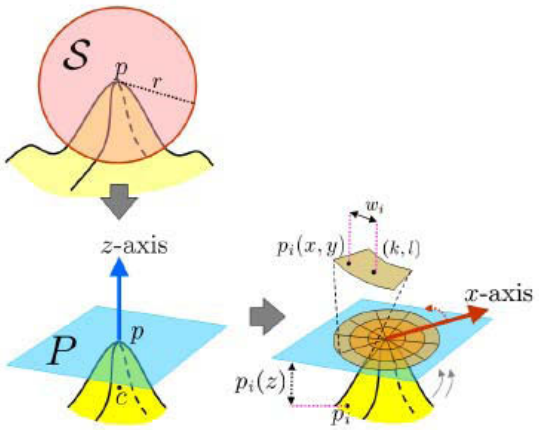


FIG. 2.17: Computation of CORS descriptor: First, throughout selecting a spherical support region, then fitting a plane to local neighborhood, and finally, selecting a reference orientation for x-axis before projecting the distances from the surface to patches. (Image taken from [Van Nguyen 11])

2.3.1 Recap

As for 2D approaches, Tab. 2.2 summarizes the most relevant 3D feature descriptors among the above-reviewed approaches:

TAB. 2.2: Recap of the most relevant 3D feature descriptors among the above-reviewed approaches

Descriptor	Reference	Evaluation domain
Spin images	Johnson [Johnson 97a]	3D-feature matching / Object recognition
3D-Fourier-Transform-based approach	Vranic et al. [Vranic 01a]	3D-feature matching / 3D-model retrieval
3D-SSD	Zaharia and Prêteux [Zaharia 01]	3D-model retrieval
Light Field Descriptor (LFD)	Chen et al. [Chen 03]	3D-feature matching / 3D-model retrieval
Spherical Harmonic Representation	Kazhdan et al. [Kazhdan 03]	3D-feature matching / 3D-model registration
3D Shape Context	Frome et al. [Frome 04]	3D-feature matching / object recognition
DESIRE	Vranic [Vranic 05]	3D-model retrieval
Tensor-based representation	Mian et al. [Mian 06]	3D-feature matching / Automatic pairwise registration of range images
3D-SIFT	Scovanner et .al [Scovanner 07]	Video action recognition / 3D imagery
Geometric scale-space-based approach	Novatnack and Nishino [Novatnack 08]	Fully automatic registration of multiple of Range Images
A compact multi-view method	Axenopoulos [Daras 09]	3D-object retrieval
CSHOT	Tombari et al. [Tombari 11]	3D-object recognition
CORS	Nguyen and Porikli [Van Nguyen 11]	3D-feature matching / 3D-object recognition

2.4 Comparison Studies of Image Feature Descriptors

In general, a competitive feature descriptor must be highly discriminative (*i.e.*, low probability of a mismatch), easy to be computed and invariant to geometric distortions such as scaling, rotation and viewpoint changes. It must also resist against lighting variations and noise. Based on these key properties, so far, few comparative studies have been conducted and designed to evaluate the performances of existing feature descriptors.

Randen and Husoy [Randen 99] proposed in their earlier work a comparison of a set of different descriptors. These include Gabor, DCT, eigenfilters, wavelet transforms, linear predictors, optimized finite impulse response filters, and Laws masks. The descriptors have been tested as various techniques of texture classification. The results show that no technique is optimal for all situations. The descriptor performance depends widely on the data at hand and the descriptor dimensionality. Despite this, Gabor filters recorded the worst results.

Carneiro and Jepson [Carneiro 02] have compared their phase-based local descriptor to the differential invariant descriptor [Huttenlocher 87]. They used in the evaluations three test images with ground truth transformations computed artificially as well as five database images. The performance criterion, *detection rate* with respect to *false positive rate*, was adopted to evaluate matching accuracies. The results display the differential invariant descriptor performing better for scale changes, whereas, the phase-based feature has best performance for illumination changes.

Ke and Sukhankar [Kuhn 05] established three type of experiments to compare their PCA-SIFT to the standard SIFT. They started by evaluating the robustness of each descriptor in synthetic images before using real images. The third experiment involved testing the descriptors in an application related to image retrieval. They adopted the performance criteria, *recall* and *1-precision*. These are often used for generating ROC-based[3] curves. The obtained curves ranked PCA-SIFT mostly better than SIFT in terms of matching accuracy and computational time for both controlled (synthetic) and real-world images.

Both descriptors proposed by Lazebnik et al. [Lazebnik 05] were tested in the context of classification. They adopted the nearest-neighbor-based matching and Earth Mover's distance (EMD) [Renninger 04]. The experiments were performed on Brodatz database [Picard 93, Liu 96] and a collection of 100 photographs of textured surfaces with different viewpoints. The performances of their descriptors were compared to differential invariant [Koenderink 87, Schmid 97] and filter banks [Baumberg 00, Cula 01, Schmid 01, Varma 02, Schaffalitzky 02] descriptors. The experiment results demon-

[3] ROC: Receiver Operating Characteristics

strated that the intensity-domain spin-images and RIFT have the best performances in terms of rotation-invariant and distinctiveness. They also placed the intensity-domain spin-images over RIFT for the texture database, whereas this latter is little better in Brodatz database.

In the seminal work of Mikolajczyk and Shmid [Mikolajczyk 05a], a number of 10 promising descriptors are investigated and compared in the context of image feature matching. These are shape context, steerable filter, PCA-SIFT, differential invariants, spin images, complex filters, moment invariants, and cross-correlation. In their evaluations, the performance criteria, *recall* and *precision* are used to plot ROC-based curves. The purpose was comparing the discriminative power of descriptors under different image deformations such as rotation, scaling, out-of-plane rotation, image blur, JPEG compression, and illumination change. Different feature matching approaches are tested. These are: nearest-neighbor, ratio-based nearest-neighbor, and threshold-based matching. Besides, different regions detectors are also used. These include Harris-Laplace [Mikolajczyk 01], Hessian-Laplace [Mikolajczyk 04], Harris-Affine [Mikolajczyk 04], and Hessian-Affine [Mikolajczyk 05c]. The performances ranked SIFT, PCA-SIFT and GLOH the most discriminative descriptors. For both textured and structured images, the higher scores are obtained with SIFT-based descriptors. As well, these perform better for image rotations and scale changes. Moreover, they recorded the larger matching accuracies for viewpoint changes. In the case of illumination changes and image blur, GLOH, PCA-SIFT and SIFT outperform the rest.

Based on Mikolajczyk's benchmark, Bay et al. [Bay 06] compared their SURF descriptor to GLOH, SIFT and PCA-SIFT. The descriptors are computed on SURF support regions. Experiments were performed for both feature matching and object recognition. Furthermore, two matching strategies are used. These are: threshold-based (similarity-threshold) and nearest-neighbor techniques. For object recognition, the performance scores placed SURF in the first rank followed by GLOH, then SIFT, and PCA-SIFT last. As well, SURF worked better in feature matching, which is much faster that remaining descriptors.

Similar to Bay et al. (*i.e.*, using Mikolajczyk's benchmark) the authors of CCH [Tola 08] pretended that the performances of CCH descriptor are better compared to those provided by SIFT.

Moreels and Perona [Moreels 07] explored the performance of different well-known detectors and descriptors for 3D-object matching. A number of different combinations detector-descriptor are evaluated in a database of 100 objects. The objects are used under 144 calibrated viewpoints and different lighting conditions. The experimental results showed that:

- Hessian-Affine detector with SIFT descriptor provides the higher robustness

against viewpoint change.

- Harris-Affine with SIFT performs best for lighting change.
- Hessian-Affine with shape context outperforms the alternatives for length changes in the camera focal.

They also noticed that no detector-descriptor combinations deal well with viewpoint changes of more than $\approx 25°$.

The spatio-temporal 3D-gradient-based [Kläser 08] descriptor is evaluated in the context of classification task. The evaluations are performed on KT [Schuldt 04], Weizmann [Blank 05], and Hollywood [Laptev 08] datasets. To compare the descriptor performance, the following descriptors are evalauted as well: Local Jets [Schuldt 04], Gradient+PCA [Wong 07], HoG [Laptev 08], and HOF [Laptev 08]. The obtained results stated that the spatio-temporal descriptor is the best at least on two out of three datasets and has the higher matching accuracy in the third.

In order to validate DAISY, Tola et al. [Tola 08, Tola 09] have conducted an elaborate experiment. They used test images (with their corresponding depth maps) of [Strecha 08]. They evaluated DAISY by comparing it to SIFT, SURF, NCC, and pixel-differencing descriptors. For dense matching task, the results obtained on blurred and low resolution web-cam images were much better with DAISY than NCC. Despite SIFT provides similar results, it is time-consuming. It is about 50 times slower than DAISY. The evaluation also highlighted the effectiveness of DAISY for the wide baseline stereo. However, for the short baseline stereo, the pixel differencing and NCC perform better. It also appeared that DAISY generates less artifacts than the other descriptors.

Feature description based on image colors are proposed recently. Therefore only few work to compare their performances are available. For object and scene recognition, Van de Sandel et al. [Van De Sande 09] analyzed the distinctiveness and invariance properties of a number of color-based descriptors.

These can be grouped in three categories. The first is related to histogram-based color approaches. These are: RGB histogram, Opponent histogram, Hue histogram [Van De Weijer 06] and RG histogram. The second category collects moment-and-moment-invariant-based color methods. The third is built around color-SIFT-based descriptors. It contains SIFT, HSV-SIFT [Bosch 06], HueSIFT [Van De Weijer 06], OpponentSIFT, W-SIFT, rgSIFT, and Transformed color SIFT.

The latter descriptors are computed for support regions obtained with Harris-Laplace detector [Mikolajczyk 01]. Besides, object appearance models are learned with the approach of Zhang et al. [Zhang 07]. The evaluation are performed on both images and videos, which include PASCAL Visual Object Classes Challenge [Everingham 07]

and NIST TRECVID 2005 [Smeaton 06, Snoek 06] datasets respectively.

In conclusion, the authors stated that object and scene recognition depends on the invariance to illumination as well as light-color. This has been criticized by Burghtouts and Geusebroek [Burghouts 09], pointing out that "no insight is gained in the effectiveness of various photometric invariants in discounting imaging effects ". They also pretend no correlation between the theoretical and the experiments.

Burghtouts and Geusebroek [Burghouts 09] proposed recent and promising framework. It aimed to compare color image descriptors using 3D-objects from ALOI dataset [Geusebroek 05]. They include in addition to their color-SIFT , CSIFT [Abdel-Hakim 06], and that proposed by Bosh et al. [Bosch 06]. These are evaluated similar to Mikolajczyk and Shmid [Mikolajczyk 05a], in which the distinctiveness is characterized by the *precision* and *recall* criteria. They used nearest-neighbor matching strategy and Harris-Affine [Mikolajczyk 04] to obtain support regions. The setup was also designed to evaluate the descriptors according to information-content properties. The results illustrated color-SIFT to be the most distinctive descriptors. They also claimed that it is robust to illumination direction, viewpoint changes, and variations of the illumination color.

More recently, an interesting comparative study is proposed by [Maji 09]. It compares performances of different descriptors in the context of object recognition. The comparison covers a set of methods, which are widely used in describing image feature. This contains SIFT, Shape-Context [Belongie 02], Spin-Image [Lazebnik 05], Image-Moment [Van Gool 96], Jet descriptors [Koenderink 87, Florack 91, Freeman 91, Florack 94], GLOH [Mikolajczyk 05a], and Geometric-Blur [Fei-Fei 06].

The methods are evaluated on the popular dataset of Caltech 101 [Fei-Fei 06]. Similar to above, the authors adopted the nearest-neighbor approach to match descriptors. Besides, the performance evaluation uses the *recall* and *1-precision* criteria to measure the distinctiveness of descriptors.

The obtained ROC-based curves showed SIFT, Geometric-Blur, and Shape-Context descriptors coming first in most reported cases. On the other hand, GLOH works poorly in spite of its high performance on Mikolajczyk's dataset [Mikolajczyk 05a] . It is also illustrated that the performance of SIFT is highly positively correlated with Shape-Context but it is highly negatively correlated with Geometric-Blur.

Chapter 3

SSC-Based Feature Description and Matching

3.1 Introduction

Our feature description and matching strategy is constructed around the four steps, early outlined in the introduction (Chapter 1). These include: `(1)` local feature description, `(2)` visual vocabulary construction, `(3)` context definition, and `(4)` feature matching.

It is worth mentioning that here we are focusing only on a subset of image data (*i.e.*, pixels or 3D points). This means we will not exploit the whole image data, but only a subset of them which has particular meaning. This subset, often called *features*, are needed to be previously selected (detected or sampled) on each image in order to be used as support elements on which the descriptors are computed during the first step.

Several techniques are available for detecting (or sampling) image features, however, the well-suited are those which are robust against noise and repeatable under different image deformations.

For instance, a number of detection methods in 2D domain, are evaluated and compared for different geometric deformations and imaging conditions by Mikolajczyk et al. [Mikolajczyk 05c]. These include: *harris-affine* and *hessian-affine* [Mikolajczyk 04], *maximally stable extremal eegions* (MSER) [Matas 04], an edge-based region detector (IBR) [Tuytelaars 99], a detector based on intensity extrema (EBR) [Tuytelaars 04], and *salient regions* [Kadir 04]. The authors concluded by claiming that no detector performs better than the others for all image types as well as for all image deformations.

Besides, other methods like SIFT [Lowe 99], SURF [Bay 06], and Harris [Harris 88a] are available when much more emphasis is up on effectiveness (in terms of computational time) than on performances.

Regarding 3D domain, a number of well-known methods for detecting 3D key points have been, recently, evaluated by Tombari et al. [Tombari 12] for 3D object detectors.

Two different categories of methods are evaluated, fixed-scale and scale-invariant detectors.

Fixed-scale detectors include those of *local surface patches* (LSP) [Johnson 99], *intrinsic shape signatures* (ISS) [Zhong 09], and *keypoint quality* (KPQ) [Mian 10].

Scale-invariant detectors contain MeshDoG [Zaharescu 09], *laplace-beltrami scale-space* (LBSS) [Unnikrishnan 08], and *keypoint quality scale invariant* (KPQ-SI) [Mian 10].

These detectors are evaluated and compared with respect to the performance, in terms of robustness, under noise, presence of clutter, occlusions, and viewpoint changes. The experimental results showed KPQ-SI, MeshDoG, and ISS to be more repeatable while

the favorite detector in terms of effectiveness is ISS.

In addition to the latter, other 3D keypoint detectors exist, and that suggested by Castellani et al. [Castellani 08], or more recently introduced in [Sun 09, Knopp 10], are examples.

Note that, we will adopt in our evaluation those of Mikolajczyk et al. [Mikolajczyk 05c] for the 2D-domain problem while that of [Castellani 08] for the 3D-domain problem. More details on these detectors are given later in Chapter 5 and Chapter 6, respectively.

3.2 Local Feature Description

Local descriptors aim at capturing local properties (*e.g.*, pixel intensity, geometric shapes and surfaces) of an image in the neighborhood of a given feature point.

In general, distinctiveness, robustness against noise, and invariance against geometric transformations are sought characteristics to obtain robust and high precision matching.

In the following, we first describe the list of local descriptors we adopt for 2D domain and then that for 3D domain. These are indicative and not exhaustive lists, opened to other local descriptors and measures.

3.2.1 2D-Domain

Three variant of SSC-based descriptor are tested. These are related to the local descriptors of:

- *SIFT* [Lowe 04]. The most popular descriptor showed to be very successful in many applications. It is a scale invariant descriptor based on the distribution of gradient magnitudes.
- *Spin Image* (SPIN) [Lazebnik 05]. The proposed *intensity-domain spin images* (abbreviated here as Spin-Image) descriptor is inspired by the standard spin images [Johnson 97a], in which the traditional coordinates are replaced by the spatial point position and brightness. The resulting descriptor is 2D histogram, which encodes the distribution of image intensities around a fixed point. SPIN has a high degree of invariance for representing affine normalized patches.
- *Cross-Correlation* (CC) [Lewis 95]. It is a basic descriptor represented by a vector of image pixels. The unnormalized CC is usually normalized using pre-computed

tables containing the integral of the image and its square over the search window. This descriptor is well suited for special effects feature tracking.

3.2.2 3D-Domain

Starting from a set of oriented points (*i.e.*, point with normal), we focus the following geometric measures to compute the descriptors for each feature point:

- *Shape Index* (si) [Petitjean 02]. The Shape Index is defined as:

$$si = -\frac{2}{\pi}\arctan\left(\frac{k_1+k_2}{k_1-k_2}\right), \quad k_1 > k_2, \tag{3.1}$$

  where k_1, k_2 are the principal curvatures of a generic vertex. The Shape Index varies in $[-1,1]$ and provides a local categorization of the shape into primitive forms such as spherical cap and cup, rut, ridge, trough, or saddle as illustrated in Fig. 3.1.

- *Beta Value* (bv). The beta value of vertex p is represented by the projection of the nearby vertex v to the normal $\mathbf{n}_p$ at p. In practice it is the distance between the surface point v and the plane identified by $\mathbf{n}_p$.

Finally, those two measurements are collected and accumulated separately onto concentric shells, centered on the feature point. This gives three-dimensional histograms $L_{si}(i,j)$ and $L_{bv}(i,j)$[1], where $i \in [1,..,N]$, $j \in [1,..,M]$ are the indices which identify the quantized geometric measures and the distance from the feature point, respectively.

3.3 Visual Vocabulary Construction

The proposed approach for learning point context is inspired from the BoW framework for textual document classification and retrieval. To this aim, a text is represented as an unordered collection of words, disregarding grammar and even word order.

The extension of BoW to visual data requires one to build a *visual vocabulary (or visual dictionary)*, *i.e.*, a set of the visual analog of words. As for [Csurka 04], the visual words are obtained by clustering local point descriptors (*i.e.*, the visual words are the cluster centroids).

[1] In practice $L_{bv}(i,j)$ is the Spin Image [Johnson 99].

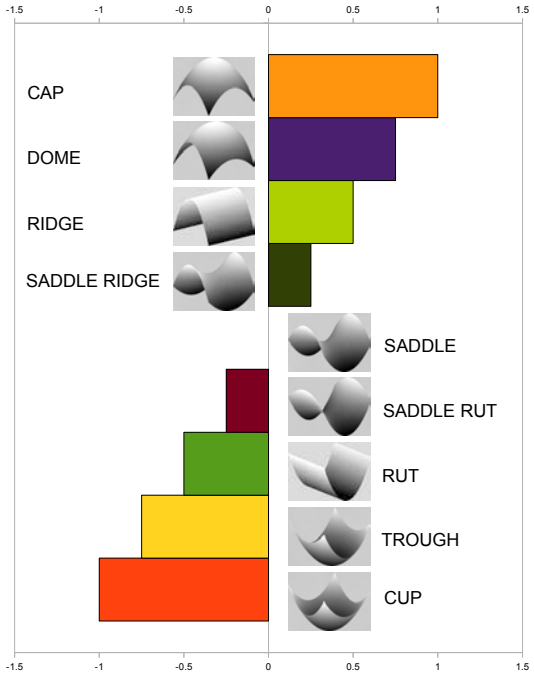
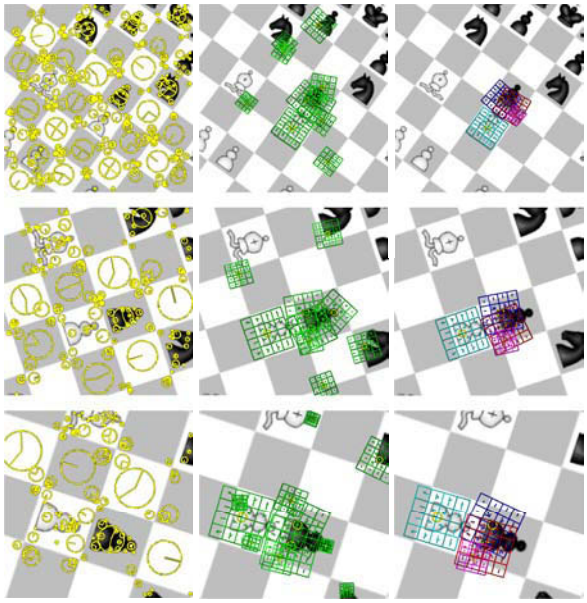


FIG. 3.1: Relationship between the shape index values and surface topology. The range of shape index values, *i.e.*, [-1, 1], can be split into sub-ranges, to provide a number of primitive forms (classes) for describing the surface local shape. For instance, in this illustration, the shape index range is divided into 8 sub-fields to obtain 9 primitive forms to describe the shape of the local surfaces.

In practice, the clustering defines a vector quantization of the whole point descriptor space, composed of all the feature points extracted from all the views (*i.e.*, the training set). In order to obtain the clustering, the k-means algorithm is employed [Duda 01]. The number of visual words is defined by fixing the parameter K.

In this fashion, each feature point can be easily classified by assigning to it the visual word associated to the closest cluster centroid. Note that in our case, as in [Csurka 04], the point classification is carried out by an unsupervised learning approach [Duda 01], but that more sophisticated classification techniques could be used.

Figure 3.2 shows the selected features, the local descriptor of each feature point detected on each image, and the feature classification on a chessboard example. Similar feature are assigned to the same visual word.



(a) (b) (c)

FIG. 3.2: An illustration of visual vocabulary construction: selected features in yellow (a), the local descriptor (in green) of each feature point (b), and the feature classification, *i.e.*, words with same color belong to the same class (c). Note that this example is based on features and descriptors selected and computed using SIFT.

3.4 Context Definition and Global Feature Description

Given one feature point, several sub-regions are defined by log-concentric shells, as for the Shape Context[2] [Belongie 02]. Therefore, a BoW representation is defined for each sub-region by counting the number of points assigned to each word. Finally, the set of BoWs composes the context definition (or SSC component). Figure 3.3 shows an example of this step.

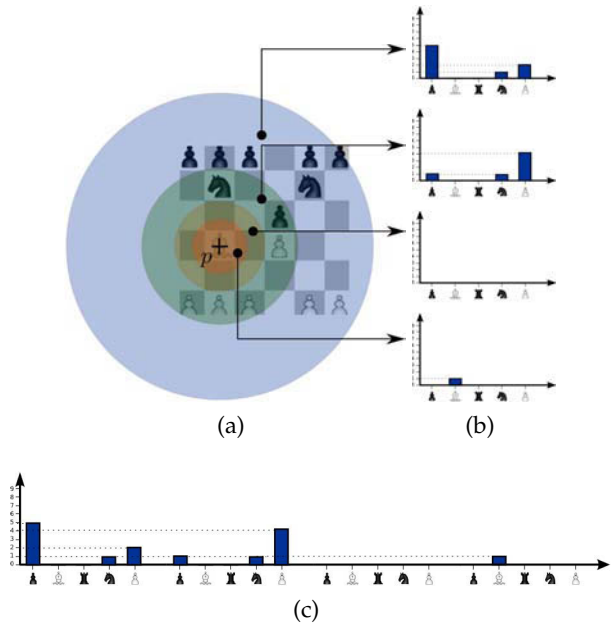


FIG. 3.3: Context definition. For clarity, we suppose that the chess pieces replace the visual words resulting from clustering algorithm. There are five visual words: *black-pawn*, *bishop*, *rook*, *knight*, and *white-pawn*. The context definition for the feature point p, is computed as follows: (a) The surrounding space of p is partitioned into log-radial concentric shells. (b) For each shell, the repeated occurrences of each word (each chess piece) are accumulated to give one 1-dimensional histogram per shell. (c) The obtained histograms are concatenated, providing thus, the context definition (or SSC component) of the descriptor computed for the feature located at the point, p.

Finally, our global feature descriptor is defined by a vector,

$$G = [L, SSC], \tag{3.2}$$

concatening the contribution of both, local descriptor and SSC component.

[2]Here the space is not split in sectors as in [Belongie 02, Frome 04] due to the instability of defining a full local reference system.

Another variant of our global feature descriptor can be built, for example, around the contribution of both shape-index and spin-image measures. Thus we obtain, G, as a vector concatenating four components:

$$G = [L_{si}, L_{bv}, SSC_{si}, SSC_{bv}] \tag{3.3}$$

3.5 Feature Matching

Matching two features is computed by comparing their respective local descriptors and their SSC components as well. For two features $\mathbf{p}_1$ and $\mathbf{p}_2$, the global similarity matrix, C, is computed as follows:

$$C(p_1, p_2) = wC_L(p_1, p_2) + (1-w)C_{SSC}(p_1, p_2), \tag{3.4}$$

where the coefficient w (ranged in $0-1$ and fixed to 0.5 in our experiments) is introduced to properly balance the contribution weights between the local descriptor and SSC component.

Besides, C_L and C_{SSC} are the similarity matrices computed by comparing the local descriptors and SSC components, L and SSC, respectively.

Note that in the 2D domain, we adopt a simple euclidean distance metric to compute C_L and C_{SSC}. However for 3D case, we use a standard metric of histograms comparison, *i.e.*, χ^2 distribution, in which:

$$C_L(p_1, p_2) = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{(H^L_{p_1}(i,j) - H^L_{p_2}(i,j))^2}{H^L_{p_1}(i,j) + H^L_{p_2}(i,j)}, \tag{3.5}$$

$$C_{SSC}(p_1, p_2) = \sum_{u=1}^{R}\sum_{v=1}^{K} g(u) \frac{(H^{SSC}_{p_1}(u,v) - H^{SSC}_{p_2}(u,v))^2}{H^{SSC}_{p_1}(u,v) + H^{SSC}_{p_2}(u,v)}. \tag{3.6}$$

The histograms $H^L_{p_1}$ and $H^{SSC}_{p_2}$ refer to the local descriptor L and SSC component of p_1 and p_2, respectively. L and SSC are of sizes of $N \times M$ and $R \times K$, respectively.

Note that, the weight function, $g(\cdot)$, has been introduced. It is related to the sub-regions (concentric shells) identified by u. The idea is to increase the influence of close regions and vice-versa. This approach is especially useful in the context of partial view (3D case) matching since furthest points are likely to be occluded. More details on $g(\cdot)$ are given in Chapter 6.

Chapter 4

Expected Performance

4.1 Introduction

As mentioned above, the proposed approach is built around combing both context and semantic information with local information. The context information is constructed around a semantic vocabulary generated for describing relationships existing between different images related to or dealing with each other. These can be, as example, images obtained from the same scene as well as range images of a same 3D full model.

The underlying idea of the semantic information is pretty straightforward. Thus, for better interaction between people of different native languages, it is necessary first to learn a common language vocabulary, like English for example. This resembles, learning the most common features, which are characterized by repetition within different images. We denominate these special features, *semantic visual features* because they express the connection existing between images. These are inspired by [Torralba 07, Matthews 02, Ullman 02, Cheng 98, Fujita 92], and also in reference to the features which are visible (seen or able to be seen) on every image.

The rationale for learning these types of relationships is that multiple images with large overlap areas will supposedly reveal an extra identical regions. We abstract the most representative set of these identical regions by a bunch of special features. This bunch constitutes the semantic features (semantic vocabulary) generated using vocabularies built around clustering algorithms like *k-means* approach [MacQueen 67, Theodoridis 06, Bradley 97].

Beside their characteristic of repeatability on different images, clusterization usually has an effect to reduce number of vocabulary words providing only the representative words and discarding less significant features. This might increase robustness against noises.

For example, the percentage of outliers can be easily reduced when the matched features are previously collected based on prior information that almost of them belong to overlap regions, which are visible on both images.

Following are our arguments that explain the advantages in adopting the SSC concept. Many of them are related to common problems in both the 2D and 3D domains, whereas certain are domain-specific problems (or applications).

4.2 Expected Performance

A reliable feature descriptor has to be discriminative and invariant. The *discriminative power* of a descriptor is its ability to distinguish between different features inside an

image. The *invariance* of a descriptor is its capacity for maintaining this discriminative power when images are subjected to different geometric transformations and imaging conditions.

Unfortunately, no single descriptor can be optimal in both, discriminative power and invariance. For instance, a descriptor naively constructed as a small square region around an image point is highly discriminative but not invariant. This is because of image transformations like rotation that point descriptors have fixed orientations, though image is rotated —in thinking about pixel to-pixel comparison of descriptors using standard Euclidean distance.

On the other side, imposing hard constraints for obtaining a full invariant descriptor can lead to a degradation of the discriminative power. In fact, a compromise or trade-off between discriminative power and invariance is often well-suited. This usually relies on the particular task , data set, and prior knowledge at hand. Such as object tracking, inside a homogeneous scene, where image transformation is approximately an *affine* only.

In general when prior information, like these latter, on the task at hand are available, by combining local and context component, it will be easy for recovering the advantages of one component in the other. For example, considered to have best discriminative power in homogeneous environment, the context descriptor can compensate for lacks in *distinctiveness* of local descriptors.

Besides, in many applications, images can be subject to major transformations, strict imaging conditions and hard viewing constraints. In particular, large scale changes, geometrical distortions and occlusions are naturally occurring. As such, to fill in the gaps when adopting a global descriptor, we need a generic complementary adaptation. For this and since in general the local descriptor performs best in presence of distortions, occlusions and for large scale changes. We think retrieving its performances into the global component, and thus we augment distinctiveness power and the invariance (robustness).

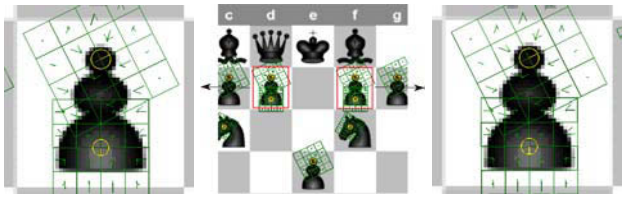
In addition, for tasks without prior knowledge, the estimation of the trade-off between distinctiveness and invariance is always difficult to be determined. In such a case, the most suitable solution is to have a configurable descriptor in which a set of tunable parameters is available to control the invariance level in order to achieve a balance between the desirable discriminative power and invariance. Thus, we could generate a suitable descriptor by varying the parameters. Allowing us to cover the full extent of trade-off from end to end and hence we could select the single descriptor appropriate for the task at hand and corresponding to the optimal trade-off.

We can also place our descriptor somewhere within this scope of approaches according to what we believe as the ideal solution to have tunable meta-parameters in rein-

forcement estimating of the distinctiveness-invariance trade-off [Varma 07] trough controlling the invariance levels. This can be done, as example, through changing the number of visual words, *i.e.*, the number of clusters.

In the following, we give some details on the expected performance of the semantic shape context approach:

- *Similar motifs and local ambiguities.* The context component, SSC, helps to reduce ambiguities occurring locally between similar regions. This has an effect to increase the discriminative power of local descriptors in scenes containing multiple similar regions in particular, as illustrated in Fig. 4.1.



(a) (b) (c)

FIG. 4.1: A typical case to demonstrate that multiple similar motifs inside an image, can result in highly ambiguous local descriptors. Though features selected inside both red boxes of (b) are spatially different, it seems by comparing (a) and (c) they have similar local presentations with SIFT descriptors. This illustration uses features obtained with SIFT detector, whereas in our experiments we adopt different detectors [Mikolajczyk 04], which will be presented later.

- *Dealing with intrinsic and extrinsic errors.* There are two major types of errors that influence the local descriptors computation: *intrinsic* and *extrinsic* errors. The intrinsic errors are related to the descriptor algorithm itself whereas the extrinsic are caused by the imprecision in detectors and errors arose from common image defects. These errors are often resulting from imperfection of image sensors and uncontrollable imaging conditions.

  Once visual words are used, these errors have less impact on local descriptor performance because the clustering-based strategy for generating the visual vocabulary helps to recover resemblance between deficient descriptors. Thus, different inaccurate descriptors computed for the same feature on different images is represented with the same visual word within all images.

  In more detail, for the intrinsic errors, though they are extracted identically on different images (*i.e.*, similar features), it happened that some of them have slight changes in their corresponding local descriptors. In particular cases, this can yield serious problems, it might increase the probability of mismatching between features, especially when errors related to the descriptor (defects related to computation stability, invariance property, etc) are accumulated.

The extrinsic errors often result from imperfection of image sensors and uncontrollable imaging conditions. For instance, some support regions can be extracted at same locations in different images, whereas they are still not covariant because of their scales and orientations which vary independently in these images.

In addition, these problems can be caused by the software and hardware materials used in the pre-preprocessing steps.

When semantic information is used, these errors have less impact on descriptor performance. This is because, the clustering-based strategy for generating the semantic vocabulary helps to recover resemblance between deficient descriptors. Thus, different inaccurate descriptors computed for the same feature on different images is represented with the same visual semantic feature within all images.

These problems seem to be easily overcome since clustering ignore slight differences that occur between same features. Thus, two identical features on two different images can be represented with a same feature even though their corresponding local descriptors are different.

- *Handling of complicated non-affine distortions.* Combing local descriptor and SSC component seems to be quite appealing for particular scenes where images present complicated non-affine distortions and non-rigid movements. These often caused by non-stationarity of objects inside images. That is to say, objects move independently during image capturing or deformation.

  For instance, the *giraffe* and *leaves* objects shown in Fig. 5.1i and Fig. 5.2f page 60, can be subjected to unpredictable motions while the scenes are under different viewpoints, *e.g.*, change of camera viewing angle. Hence, it becomes difficult for matching features within images obtained from these scenes assuming affine warps (or even rigid transformations). By allowing non-affine image transformations, the local descriptors may fail to address the matching and thus the registration problem.

- *Invariant to translation and rotation.* It is intuitive to consider the invariance of the SSC component to translation, since all histograms are built relative to points on the image.

  The SSC component, such as it is, is inherently rotation invariant because the accumulated occurrences of each visual word inside each concentric shell are nearly unchanged under rotation. The resulting histograms are notably insensitive to arbitrary rotations applied to images.

  Furthermore, the rotational invariance is reinforced with the rotation-invariant local descriptor, like SIFT for example. This is because, both components are contributed. In contrast with the proposed approach, all the state-of-the-art

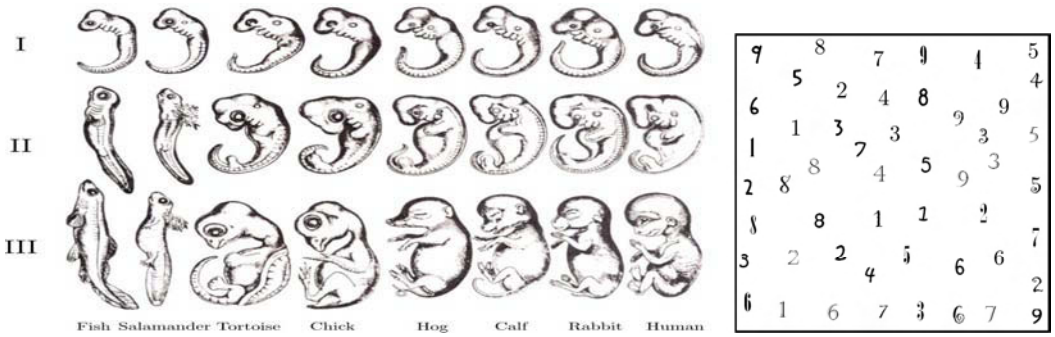
approaches, like Shape Context [**?**] for example, requires previously determining an accurate relative orientation in order to have rotational invariance. This usually involves additional computations yielding to decreasing in the descriptor accuracy and an excess in the complexity time.

- *Occlusion handling.* Occluded points for 3D models can yield serious problems for many feature matching related tasks. These points usually visible on an image but not on others, can cause a decrease in the number of correct matches, *i.e.*, even though one point is detected on the first image, its potential correspondence on the other image is still hidden. Since the SSC component is constructed around points mostly visible on different images, it can be solution to deal with occlusions appearing in scenes of large occluded parts, like 3D partial views.

- *Imperfections and discontinuities of surface.* The problems when adopting a local approach for describing 3D model features are caused by the errors related to surface imperfections. For instance, the presence of multiple discontinuities, or *holes*, on surface is typical and common defects for many scanned 3D models. These defects are, in general, due to noise and systematic errors arising during creating range surface.

- *Object shape similarities.* Many machine vision applications depend heavily on object recognition and classification. For objects like those of Fig. 4.2a and Fig. 4.2b, the similarities across embryonic (*e.g.*, the first row in Fig. 4.2a) and typewritten (handwritten) digits are evident. Thus, a number of recognition and classification related tasks might be performed unsuccessfully once shape information are adopted (global, or context, approaches). However, when it is added to the shape information, local information might improve the recognition and classification accuracies. This is because, for point-to-point comparison, these objects are different while they are quite similar in terms of shape-to-shape comparison.

  The proposed method not only compensates for errors arose from one category of approaches into other (*i.e.*, from local into global and vice versa), it is also designed to make semantic-context component working in collaboration with the local descriptor. That is, these semantic-context and local components, contribute optimally in improving feature description. In the sense that the context information is built around semantic features, which are in turn generated based on local descriptor component. This allows the contribution of the context component more efficient since a same set of semantic features are used for describing all images. It helps, hence, to abstract the features mostly visible on all images.



(a) (b)

FIG. 4.2: Example of objects with similar shapes. (a) Haeckel's drawing [Richardson 02], a copy of The Romanes 1892. This figure shows vertebrate embryos at different stages of development for (from left to right) *Fish*, *Salamander*, *Tortoise*, *Chick*, *Hog*, *Calf*, *Rabbit* and *Human*. Taken out of context (only for instructive purpose), the embryonic similarity is evident across the images of the first row of (a). Thus, many machine vision tasks depend heavily on object recognition and classification, can be performed unsuccessfully. (b) A set of different typewritten and handwritten digits. In terms of shape-to-shape comparison, there are many similarities between digits, but for point-to-point comparison, these digits are quite different.

Chapter 5

Experimental Results – 2D Domain: Image Feature Matching

5.1 Introduction

In this Chapter we focus on the 2D domain, for which the performance sought of a 2D feature descriptor is to be *discriminative* and *invariant* against different geometric transformations and imaging conditions.

The SSC approach is tested in 2D domain by addressing a feature matching problem. In this case, we consider 2D images selected from a well known dataset, used habitually to evaluate and compare the performances of 2D image feature descriptors.

Besides, we also include another dataset intentionally created from images containing multiple similar motifs and depicting complicated non-affine distortions, *i.e.*, slight non-affine transformations.

The performance of the descriptors is evaluated according to the discriminative power and invariance criteria using a well-known standard benchmark.

This Chapter is organized as follows: In Section 5.2, we outline the evaluation scheme while presenting data set, performance criteria, descriptors, and matching strategies. In Section 5.3, we show the evaluation results. The conclusion is given in Section 5.4.

5.2 Experimental Setup

The performance of SSC-based descriptor is evaluated using the standard benchmark of *Mikolajczyk* [Mikolajczyk 05a] available on-line [1] . This includes supports of programs and dataset for evaluating and comparing descriptor performances in the context of image feature matching.

5.2.1 Computation of SSC Component

To compute SSC components, we have coded a basic `c/c++` implementation, which is compiled on `Intel(R)Core(TM)2 Duo CPU P8700 @2.53GHz` machine model, running on `Linux x86-64` architecture.

[1] http://www.robots.ox.ac.uk/~vgg/research/affine/

5.2.2 Data Sets

The experiments are performed on two different data sets. The first we intentionally selected from homogeneous environment of multiple similar regions and complicated non-affine distortions, while the second is that of Mikolajczyk available on-line. [2]

For the first set, *i.e.*, of homogeneous environment and complicated non-affine distortions, images are obtained from both the *structured* and *textured* scenes, which reflects natural environments and real operating conditions of our SSC-based descriptor, expecting thus to provide better performance than other descriptors. Fig. 5.1 displays an image example for each scene in this data set.

The scenes are selected from video sequences such that its corresponding images (frames) are related by planar projective transformations, *i.e.*, *homography*. This is one-to-one mapping between two images, which describes the image motion between two frames when *(i)* the camera motion is pure rotation, or *(ii)* the camera is viewing a planar scene [Prince 02].

Each scene of this dataset contains a set of images depicting gradual increases in specific geometric transformations. These include rotation, scale change, and viewpoint change. The ground truth transformation that maps each image to its reference is computed with a similar approach of [Mikolajczyk 05a].

Thus, we start by selecting manually a reduced number of point correspondences between the reference and each image. Based on these correspondences, approximate homographies are estimated, which are then used to align each image to its reference.

Next, in order to compute precise homographies, a small-baseline-based robust method for computing accurate residual homographies is applied between the reference and the aligned images. The accurate homographies which map images to their reference are obtained as compositions of the approximate and residual homographies.

The second data set is the standard dataset of Mikolajczyk. This contains images which are subjected to different geometric transformations and imaging conditions. In terms of geometric transformations, it includes rotation, scaling, and viewpoint change. For imaging conditions, there are, image blur, illumination change, and JPEG Compression. More details on this data set are available on-line. [2] Fig. 5.2 shows image samples from this data set.

[2] http://www.robots.ox.ac.uk/~vgg/data/data-aff.html

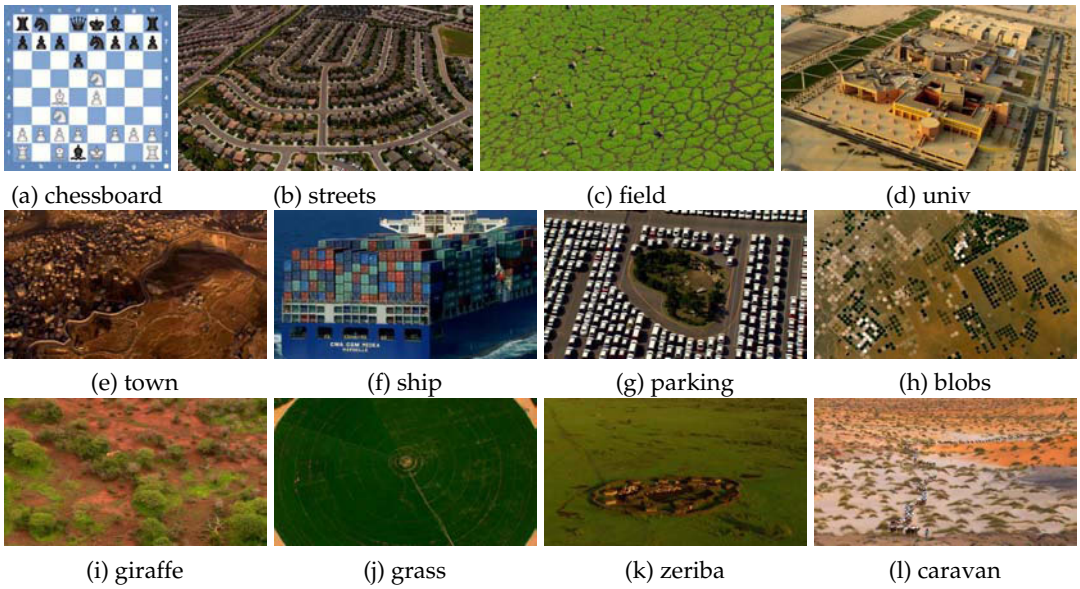
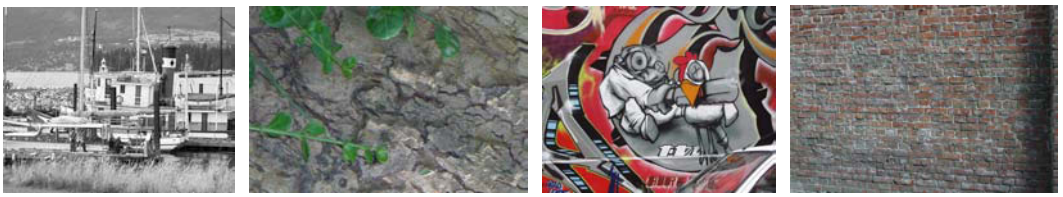
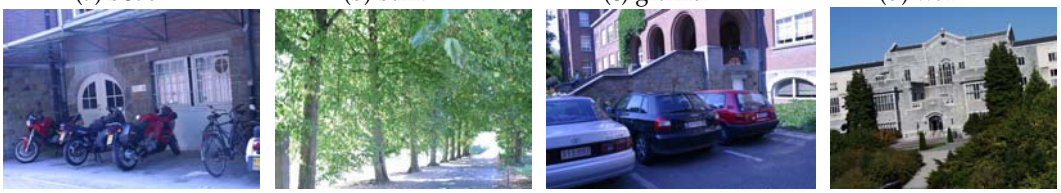


FIG. 5.1: The data set we have created for evaluating the descriptors inside challenging environments. This is intentionally generated from structured and textured scenes of homogeneous environment and complicated non-affine distortions. It reflects geometric transformations related to (a)(b)(d) structured image rotations, (c) textured image rotations, (f)(g)(h) structured image scaling, (e)(i)(j) textured image scaling, and (k)(l) viewpoint changes in textured environments.



(a) boat (b) bark (c) graffiti (d) wall



(e) bikes (f) trees (g) cars (h) ubc

FIG. 5.2: The standard dataset we have used in our experiments. This is the well-known dataset of Mikolajczyk. It contains structured and textured images subjected to (a)(b) combined rotation-scale, (c)(d) viewpoint changes, (e)(f) image blur, (g) illumination change, and (h) JPEG compression.

5.2.3 Evaluation Criteria

The descriptors performances are evaluated according to both the discriminative power and invariance criteria.

5.2.3.1 Discriminative power

It measures the ability of a descriptor to distinguish between different features in an image. Similar to [Burghouts 09, Maji 09, Bay 08, Mikolajczyk 05a], we evaluate the discriminative power through ROC-Based curves, which show *recall* as function of *1-precision*. We use 1−precision instead of precision in order to be compatible with the standard ROC graph [Fawcett 04].

The recall score corresponds to the ratio of the number of correct matches to the number of correspondences. Whereas precision is the ratio of number of correct matches to the total number of matches. In an equivalent manner, the 1−precision is the ratio of the number of false matches to the total number of matches. The latter scores are computed based on the following formulas:

$$recall = \frac{\#correct\_matches}{\#correspondences}, \quad 1 - precision = \frac{\#total\_matches - \#correct\_matches}{\#total\_matches} \tag{5.1}$$

Here, $\#correct\_matches$ is determined based on the following conditions:

- The Features are matched in the descriptor space. A pair of features in two different images is qualified as correct match if the matching of their corresponding descriptors is verified with the related matching strategy, *e.g.*, in the case of threshold-based matching, two features constitute a correct match if the distance between their descriptors is below a threshold.
- The match identified as correct in the descriptor space must be also correct in the correspondence space, *i.e.*, belongs to the set of correspondences.

The $\#correspondences$ is calculated using the ground truth transformations (homographies) and region overlaps between image features. Briefly, a pair of features in two different images are supposed to be in correspondence if their normalized region overlap error, expressed as a percent fraction (%), is below a threshold (fixed to 50% in our experiments). Explicit details can be found in [Mikolajczyk 05a, Mikolajczyk 05b].

The *cut-off points* of ROC curves are obtained by varying the value of $\#total\_matches$ in a fixed range. In turn, these points correspond to different thresholds of the matching strategy at hand.

Since the ROC curve is considered to depict the relative *trade-off* between the recall (profit) and 1-precision (cost) scores, an ideal discriminative descriptor has its curve passing through the upper left corner, *i.e.*, *1-precision=0* and *recall=1*. Therefore, the closer the ROC curve is the upper left corner, the higher the overall discriminative power of the descriptor. This also means, the larger the area under the ROC curve, the higher the discriminative power of the descriptor.

There are particular points on ROC curve. The origin (1-precision=0, recall=0), with such a descriptor, no correct matches have been correctly identified. This point represents a case when the descriptor never produced a correct match. It may occur if the matching threshold is too low for the descriptor.

The upper right point (1-precision=1, recall=1) represents a case where all correct matches are correctly identified (recall=1) while no false matches are correctly identified (1-precision=1). In this case, the descriptor recognizes all correct matches but it has high probability to incorrectly identify false matches as correct.

The upper left point (1-precision=0, recall=1) represents the ideal point in which all correct matches are correctly identified and all false matches as well. Therefore, we should always trust the descriptor, since the probability to add false matches to the set of possible correct matches becomes null.

In general, the left-hand side of an ROC graph is more interesting, since in many tasks the matching is dominated by large number of false correspondences.

5.2.3.2 Invariance

It evaluates the ability of the descriptor performance (*e.g.*, discriminative power) to remain high and unchanged when an image is strongly altered by particular geometrical transformations and imaging conditions. The invariance, or robustness, measures the *constancy* of the descriptor performance under gradual increase in image degradation.

Unlike other approaches, we evaluate the change in both recall and precision scores. This is more relevant than considering the invariance with respect to the recall score only.

5.2.4 Descriptors

The performance of three variants of SSC-based descriptors, SIFT-Based-SSC, SPIN-Based-SSC and CC-Based-SSC, are evaluated and compared to ten state-of-the-art approaches described in Chapter 2.

These include the local descriptors of SIFT [Lowe 04], spin images (SPIN) [Lazebnik 05], complex filters (CF) [Schaffalitzky 02], differential invariants (KOEN) [Koenderink 87], steerable filters (JLA) [Freeman 91], moment invariant (MOM)[Van Gool 96], normalized cross-correlation (CC) [Lewis 95], GLOH [Mikolajczyk 05a], and PCA-SIFT [Ke 04]. In addition, we incorporate the contextual approach of Shape Context (SC) [Belongie 02].

5.2.5 Matching Strategies

Image feature matching performance depends heavily on the matching approach. For this reason, the descriptor performances are evaluated with different matching techniques. These are *nearest-neighbor*, *distance-ratio-based* nearest-neighbor, and *threshold-based* matching.

5.2.5.1 Nearest-neighbor

The nearest-neighbor is mostly correct (*i.e.*, with higher precisions), since it selects only one match (*i.e.*, the nearest neighbor) below a threshold while discarding all the rest. This explains why their corresponding ROC curves are mostly nearby the left-side region of ROC graph.

5.2.5.2 Distance-ratio-based nearest-neighbor

The distance-ratio-based nearest-neighbor is similar to the nearest-neighbor, unless the threshold is directly proportional to the ratio between distances to the first and second nearest-neighbors. For instance, feature f_i on an image is matched correctly (*i.e.*, considered as a correct match) to its first nearest-neighbor f_j on another image, if only if the ratio, $dist(f_i, f_j)/dist(f_i, f_k)$, is below certain threshold and by taking into account that f_k is the second nearest-neighbor of f_i. Here $dist$ is a metric (*e.g.*, euclidean, mahalanobis, correlation, *etc.*) to measure the distance between two feature descriptors.

In general, distance-ratio-based nearest-neighbor matching is less accurate than nearest-neighbor especially for images of multiple similar regions, and thus, it is not will suited on this type of images because of many false matches can be introduced.

5.2.5.3 Threshold-based matching

For threshold-based matching, two features constitute a correct match if the distance between their descriptors is below a threshold. Therefore the feature can obtain lot of matches and many of them can be false which leads to low precision scores. This explains why its corresponding ROC curves are often extended over the precision range. This approach is useful when we envisage more efficiency than performance, particularly for dense matching, *i.e.*, within large feature databases.

5.2.6 Feature Detectors

In order to evaluate the impact of the errors arise from a lack of *invariance* and *accuracy* of region detectors (*i.e.*, here features as obtained as regions and used as supports of descriptors), the descriptors are computed on regions obtained with different detectors. The detector errors are often related to the inaccuracy in size, spatial position, and orientation angle of the region as well as the insufficient robustness (invariance) of the detector under particular image transformations.

The region detectors we select are those of *Mikolajczyk* [Mikolajczyk 01, Mikolajczyk 04, Mikolajczyk 05b], mainly constructed around the standard Harris corners detector [Harris 88b]. Following are some details about these detectors.

5.2.6.1 Harris-Laplace

It is a similarity invariant detector, *i.e.*,the invariance is against translation, image rotation and scale change only. The algorithm is performed in two steps. First, for each scale of a preselected range, a scaled-adapted Harris matrix is computed. Based on these matrices, a list of interest point sets is extracted (similar to the standard Harris approach). Thus, each item (or interest point) in the constructed list is defined by its spatial-location and scale coordinates. Next, an exhaustive search in the scale space is applied on the whole list. The purpose is to select the items of scale with local maximums of the Laplacian-of-Gaussian.

5.2.6.2 Hessian-Laplace

It is similar to Harris-Laplace –similarity invariant detector– except the interest point are extracted based on Hessian matrix instead of Harris matrix (*i.e.*, covariance matrix).

5.2.6.3 Harris-Affine

It is an affine invariant detector, *i.e.*, the invariance is with respect to translation, image rotation, scale change, and image shearing. The algorithm is a two-steps process. The first involves using Harris-Laplace detector to determine the spatial location and scale of the regions. The second is a refinement step based on an affine adaptation algorithm. The purpose is to improve the region shape in order to be robust against image affine deformations.

5.2.6.4 Hessian-Affine

The approach is similar to Harris-Affine (*i.e.*, invariant to affine transformation as well as based on shape adaptation algorithm), unless it uses Hessian-Laplace instead of Harris-Laplace for computing the spatial location and scale of the regions.

Because of their blob-like shapes, the regions obtained by Hessian-Laplace and Hessian-Affine approaches are more conservative of information. Furthermore, they are more

accurate than those obtained with Harris-Laplace and Harris-Affine approaches since the scale selection in Hessian-Laplace is more precise than in Harris-Laplace.

5.3 Results and Discussion

In this section, we compare the performance of variants of SSC-based descriptors to those obtained with ten of state-of-the-art approaches listed in Section 5.2.4 and well described in Chapter 2. As for SSC-based descriptors, we include SIFT-Based-SSC, SPIN-Based-SSC, and CC-Based-SSC.

The comparisons are performed for data sets of 20 scenes presented in Section 5.2.2. These are selected to reflect different image alterations, including geometric transformations and imaging conditions. The images composed our created data set, *i.e.*, shown in Fig. 5.1, are mainly obtained from homogeneous and complicated environments , expecting thus that the previous variants of SSC descriptors will perform much better than the other descriptors.

To inspect how the performances of different descriptors are influenced by extrinsic errors arise from lacks of accuracies of both the detector[3] and matching strategy, the descriptors are compared for different detectors and matching techniques. To this end, we include Harris-Laplace, Hessian-Laplace, Harris-Affine, and Hessian-Affine region detectors. More details on these detectors are given in Section 5.2.6. According to matching approaches, we adopt nearest-neighbor, distance-ratio-based nearest-neighbor, and threshold-based matching techniques. These are presented in Section 5.2.5.

Mostly, we will adopt the nearest-neighbor and threshold-based as the main matching techniques. We select nearest-neighbor because it is well-suited for image feature matching since it is usually correct with high precision scores. Whereas for large number of features, it is difficult to use. Hence, it becomes adequate to apply threshold-based matching.

The performances are evaluated based on both the discriminative power and invariance criteria, previously described in Section 5.2.3. Briefly, the discriminative power measures the descriptor distinctiveness. It is presented by ROC-based curve, which plots *recall* as function of *1-precision* scores. The invariance (*i.e.*, robustness or stability) evaluates the *constancy* of descriptor performance under gradual increase in image alterations. The descriptor invariance is measured with respect to both the recall and precision scores for different geometric transformations and imaging conditions. Here, the invariance is also evaluated with respect to the precision and not only according

[3]That is to say, those related to features and here are obtained as support regions

to the recall as seen in the literature. We believe this is more revealing than using the recall only.

Since the number of correspondences (*i.e.*, #correspondences as described in Section 5.2.3), which is used to find the number of correct matches, entirely relies on the amount of overlap error and the ground truth transformations (homographies), we will evaluate how well the region overlap errors and the descriptor performances are correlated. For this purpose, we will measure the degradation in the recall and precision scores according to gradual increases in the overlap error. This error is fixed to 50% in the other evaluations.

It is worth noting that the left-side of the ROC graph, which corresponds to the range of high precisions, is the most relevant region for comparing the discriminative power of descriptors. Therefore, the focus upon this region will be more important.

In addition to all of the above, we present across Section 5.3.2, some evaluation results showing the effect of some parameters on the performance of SSC-based descriptors. Though the focus of this work is more on performance than on effectiveness, we also include in Section 5.3.10, an overview over of the computation time reported with our approach.

Figure representation: for the purposes of clarity, the following conventions will be adopted for figures and curves:

- For each figure, solid lines of red color are used to plot the SSC descriptor curves.
- In each curve, the same marker symbol is used for both SSC descriptor and its related local descriptor.
- The SSC descriptors are labeled with bold text in the bottom of each legend box.

5.3.1 Results Overview

Before going into details, we include in this section an overview of the results and most relevant observations noted during our experiments.

First, we briefly present in Fig. 5.3 a preview of the evaluations results, which summarizes the performance, in terms of recall percentages, of different evaluated descriptors with respect to different image deformations. It is easy to observe how well SIFT-Based-SSC outperforms the other descriptors on both the structured and textured scenes as well as for all the types of image deformations.

Overall, the experiments show at first sight, that the SSC descriptors perform best within images taken from homogeneous scenes while maintaining higher discriminative power and invariance. Besides, the performances of other descriptors are disappointing with homogeneous scenes, and the most successful local and global descriptor, like SIFT and Shape-Context respectively, are deficient in performance.

It seems that for certain textured scenes with large number of similar regions, the local descriptors turn to be unsuitable, as illustrated in Fig. 5.17a and Fig. 5.17b. Even though one would expect that for these scenes, we could achieve the best performance with contextual approaches (*e.g.*, shape-context), they turn out to be less successful compared to our SSC technique.

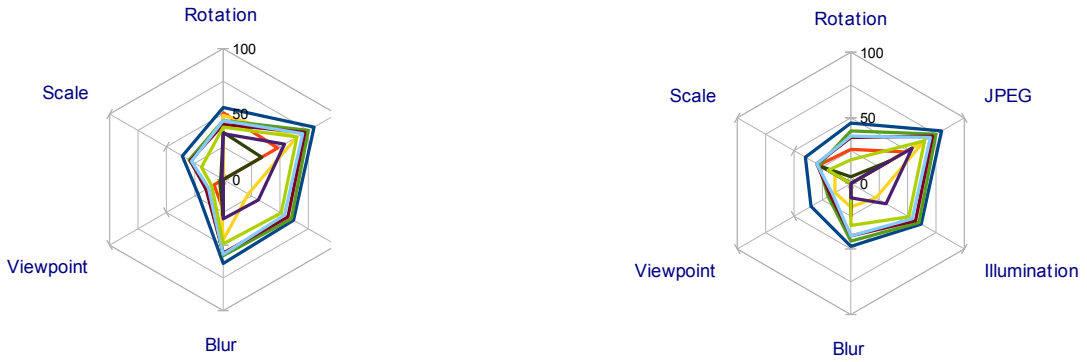
This supports our claim that adding the semantic information, better discriminative power and invariance could be obtained. Contrary to these scenes, the SSC approach performs less in images containing a reduced number of similar motifs but still better than other methods. Overall and according to the experimental setup, we retained the following points:

Weighting factor. A heuristic evaluation showed that a value of the weighting factor, w, between 0.5 and 0.6 gives good performances.

Data sets. With respect to the scene type, the SSC descriptors perform better in scenes presenting a large number of similar regions. We figured out that, higher the number of similar regions, the higher overall discriminative power. This can be checked through observing the discriminative power of the descriptors for scenes with large similar motifs (*e.g.*, Fig. 5.1a and Fig. 5.1c) and those with less number of similar motifs (*e.g.*, Fig. 5.1b). This is clarified by comparing the ROC curves of Figs. 5.6a and 5.14a to those of Fig. 5.10a.

Image alterations. With respect to different image degradations, the approach has better performance against geometric transformation than against imaging conditions.

For example, Fig. 5.3 illustrates a better increasing in discriminative power of SIFT under rotation, scaling, and viewpoint change, when adding SSC information. This is observed through comparing the gaps between the blue (SIFT-Based SSC) and red (SIFT) markers of, rotation, scale, and viewpoint change to those of blur, illumination, and JPEG compression.



(a) Structure

(b) Textured scenes

FIG. 5.3: Experimental results preview. This compares the descriptor performances for different types of image deformations, in terms of recall percentages. The recall percentages are computed for precision values varying between 80% and 95%.
For the results shown in (a), we have included the scenes of `chessboard` (image rotation) , `boat` (scale change), `graffiti` (viewpoint change), `bikes` (image blur), `cars` (illumination change), and `ubc` (JPEG compression).
For those of (b), we have used the scenes of `field` (image rotation), `giraffe` (scale change), `zeriba` (viewpoint change), `trees` (image blur), `cars` (illumination change), and `ubc` (JPEG compression). Note that the both sub figures share the same legend.

Detectors. For the different types of tested detectors (*i.e.*, support regions), the best performances in both the discriminative power and invariance are recorded with our SSC approaches. The gap in the precision between our descriptors and the other is more apparent for Hessian-based regions (*i.e.*, Hessian-Laplace and Hessian-Affine) as shown in Fig. 5.7a and Fig. 5.9a. This is due to the fact that Hessian-based detector is more accurate than Harris-based detector.

Matching strategies. Regarding the matching approach, the performances of all the descriptors are better with the nearest-neighbor than those for the distance-ratio-based nearest neighbor and threshold-based matching techniques. This is observed, as example, across comparing the curves of Fig. 5.7a to those of Fig. 5.6b and Fig. 5.7b.

Mainly this is caused by the fact that the nearest-neighbor is mostly correct. Which is not the case for the two other matching methods, since it may occur that one feature may have a number of different matches even though only one of them is correct. Thus, there are more number of false matches.

For instance, with the distance-ratio-based nearest-neighbor the number of correct matches is minimized in the sense that the matching threshold is set as function of the first and second nearest neighbor (see Section 5.2.5).

Computational time. To find out how the computational time influences the performance, especially, when varying the parameters values of the SSC component, we have measured the computation time of SIFT-Based-SSC component for different numbers of semantic words (clusters) and images.

The results showed that when setting the number of clusters, $k = 25$ as well as that of images to 2, the average time (*i.e.*, wall clock time) required to compute each SSC component (*i.e.*, per feature) is 1.967 ms, whereas we reported 4.801 ms with SIFT taken alone.

This means SSC component is less time consuming. It is approximately 41% of SIFT computation time. This is very motivating since the focus here, is more upon the performance than efficiency. Therefore, the constraint of the computational complexity, engendered by the additive computation time of SSC component, can be improved when the emphasis upon it, will be more important.

In the following, we first present and discuss the results related to evaluate the performance of our proposed SSC descriptor when some of its relevant parameters are varied. Next, we illustrate the experimental results obtained on evaluating and comparing the performances of the descriptors for different geometric transformations and imaging conditions. Lastly, we briefly discuss the effectiveness of our proposed SSC descriptor

while providing some of reported computation times.

5.3.2 Effect of Parameter Setting

Our feature description and matching approach has three adjustable parameters:

- the weighting factor, w, between local and context components.
- the number of visual vocabulary, k.
- the number of log-concentric shells, s, around each feature location.

In addition, two other parameters related to the radii of the inner and outer shells, are set. The robustness of SSC components against scale changes depends on the values of these parameters.

The heuristic evaluation illustrated in Fig. 5.4 suggests that the reasonable range of w is $0.4-0.7$. This is fixed to 0.5 in our experiments.

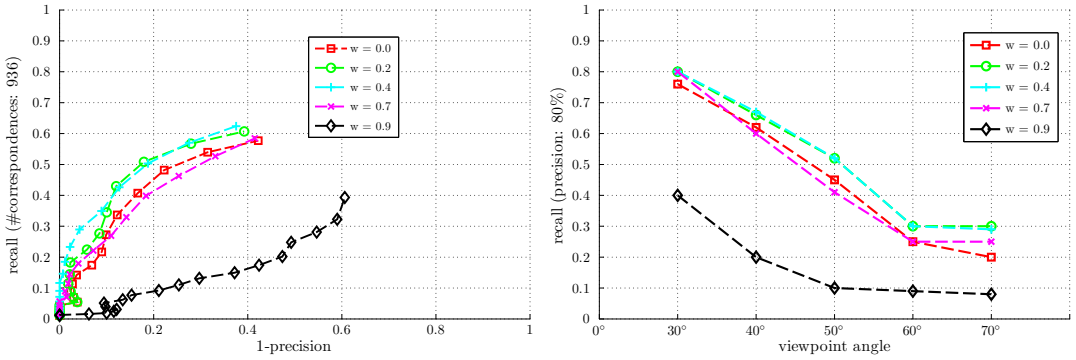
Though a low value of k is quite sufficient for homogeneous scenes (*e.g.*, textured images), a high value is probably required when dealing with structured scenes. In this spirit, we set $k = 25$ in the all experiments.

A similar investigation to the above is conducted to determine the effect of varying of the number of log-concentric shells, s, on the performance of SIFT-based-SSC. This showed that a value around, $s = 12$ (the number of concentric shells used to compute SSC components), is reasonably quite enough to ensure good performances.

To minimize the effect of the scale on the SSC robustness, we adopt an approach similar to [Belongie 02]. The idea is setting the inner and outer radii to 1/8 and 2 respectively, after normalizing (*e.g.*, by the mean) of pairwise euclidean distances of all the image feature points.

5.3.3 Image Rotation

In this experiment, we evaluate each descriptor performance under image rotation for structured and textured scenes. For structured, we use the scenes of `chessboard` and `streets` of Figs. 5.1a and 5.1b, respectively. These reflect images of large and medium number of similar motifs. According to the textured scene, we use images of `field`, shown in Fig. 5.1c.



(a) Discriminative power (b) Invariance

FIG. 5.4: Effect of weighting factor, w, on (a) discriminative power and (b) invariance. The results are with respect to viewpoint changes using the textured scene of `zeriba`. The discriminative power is evaluated according to a viewpoint angle of 50°.

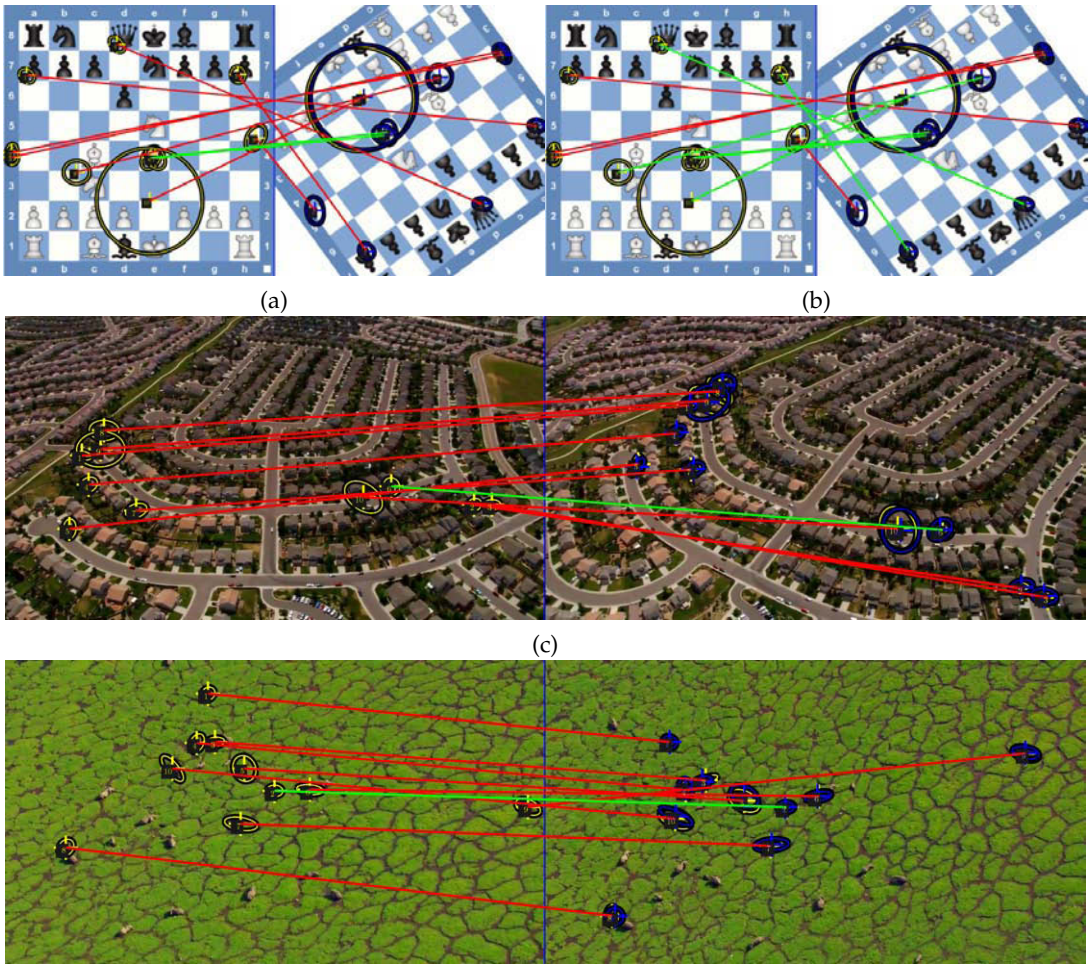
The scene `chessboard` contains 4 images presenting different rotation angles relative to the reference frame shown in the left side of Fig. 5.5a. The rotation angle varies in the range of $36° - 144°$. The scene `streets` is composed of 6 images with different rotations from the reference image of the left side of Fig. 5.5c. The rotation angle varies approximately between 10° and 40°. The third scene, `field`, contains 6 images subjected to affine transformations dominated by rotation. The rotation angles are approximately between 5° and 25°.

The descriptor performances are evaluated according the discriminative power and invariance. For the discriminative power, the evaluation is performed with the three scenes, and each descriptor is computed for the four region detectors and matched with different matching techniques. We adopt the nearest-neighbor and threshold-based matching strategies in almost all evaluations.

For the invariance, we use `chessbord` scene, and the descriptors are computed for Harris-Affine and Hessian-Affine regions and then matched with the nearest-neighbor and threshold-based matching techniques.

An example illustrating the resulting nearest-neighbor matching using a sample of descriptors computed for different region detectors is shown in Fig. 5.5. The detected regions are colored yellow, while their correspondences transformed from the reference (*i.e.*, image in the left of each sub figure) to the second (*i.e.*, image in the right of each sub figure) using the ground truth are in blue.

The correspondences computed based on ground truths and overlap errors are highlighted with blue lines, whereas matches identified as correct using descriptors are highlighted with green lines. For instance, in Figs. 5.5a and 5.5b, we use PCA and



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

FIG. 5.5: An example of nearest-neighbor matching using a sample of descriptors computed for different region detectors. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors and overlap errors (ground truths) are highlighted with green and red lines, respectively.
For example, in (a) and (b), we use PCA and SPIN-based-SSC, both computed for regions selected with Harris-Laplace detector. However, (c) and (d), are with SIFT-based-SSC computed for Harris-Laplace and Harris-Affine regions respectively.

SPIN-based-SSC approaches, both computed for regions selected with Harris-Laplace detector.

However for Figs. 5.5c and 5.5d, we adopt SIFT-based-SSC computed on Harris-Laplace and Harris-Affine regions, respectively.

5.3.3.1 Discriminative power

In this experiment, the discriminative power of the descriptors is evaluated for `chessboard`, `streets`, and `field` scenes. For each scene, the descriptors are computed on different support regions and then matched with different matching approaches as well. The obtained results are highlighted in the figures range from Fig. 5.8 to Fig. 5.17.

Overall, it seems SIFT-Based-SSC recorded the best scores over most experimental settings. This means for different scenes, support regions, and matching algorithms. We notice that the performances in discriminative power of SPIN and the basic CC descriptors are considerably improved when adding semantic context information. Thus, the resulting CC-based-SSC becomes more effective than competitive descriptors like SIFT as we can see in Figs. 5.7a, 5.8a, and 5.9a.

We also observe that the best performance of SIFT-Based-SSC over all scenes is obtained on the textured images of `field`, which contains a large number of similar motifs. This is illustrated across Figs. 5.14a–5.17b

In detail, for the scene of `chessboard`, which is considered as structured scene containing the larger number of similar regions, we observe that all SSC-based descriptors outperform the other descriptors. Thus their ROC curves are mostly the closest to the left side.

The gap in discriminative power between SSCs and the others is more apparent for the descriptors computed on `hessian-laplace` and `hessian-affine` regions and matched with the `nearest-neighbor`. This is because the Hessian-based detector is more precise since they select more accurate regions than Harris-based detectors. In addition, this is because of the `nearest-neighbor`, which is mostly correct.

This statement is reflected in Figs. 5.7a and 5.9a, in which the ROC curves of SSCs are closer to the upper left side than others (*i.e.*, the closer the ROC curve in the upper left corner, the higher the overall discriminative power).

It is worth noting that SIFT-Based-SSC, SPIN-Based-SSC, and CC-Based-SSC are always ranked in the top spots for all detectors and matching methods. In particular, the CC-Based-SSC which is constructed around the basic cross-correlation approach shows to be more competitive than SIFT and GLOH.

For the `streets`, which is regarded as a structured scene containing a moderate number of similar regions, the performance of SSCs is less than those obtained with the `chessboard`. However, SIFT-Based-SSC still outperforms all descriptors, even though it recorded a noticeable decrease in discriminative power, as observed in Figs. 5.10 up 5.13.

The descriptor ranking is no longer conserved except for SIFT-Based-SSC which still in the top spot. Although they lost their best ranking in favor of SIFT, GLOH and SC, the SPIN-Based-SSC and CC-Based-SSC still perform better than SPIN and CC. In particular, CC which is highly improved when adding the SSC information.

Concerning the `field` scene, which is composed of highly textured images, SIFT-Based-SSC descriptor exhibits higher discriminative powers. The gap in the discriminative power between SIFT-Based-SSC and the other descriptors is too large especially inside the ROC region of high precisions, *i.e.*, *1-precision* considerably small.

This range of precisions corresponds to the most revealing region in the ROC graph for evaluating the discriminative power of a descriptor. We observe upon Figs. 5.14–5.17 that certain descriptors turn out to be completely obsolete for the high precision values, while SSCs provide high recall scores.

Moreover, in general, the global (or context) approaches as Shape-Context (SC) have been shown to perform always better than the local techniques for the textured scene like `field`. However through inspecting Figs. 5.14a, 5.15a, 5.16a, and 5.17a, it appears without any doubt that the SC is completly outperformed by the SSC descriptors.

Thus, we remark that for the high precision range while the local descriptors like SIFT produces recall scores nearby zeros, the SIFT-based-SSC is providing an extremely high score (≈ 0.35). This also means that certain local descriptors, like SIFT, fail completely to find correspondences at high precisions, whereas SIFT-based-SSC find very high number of correspondences at high precisions.

The Tab. 5.1 is given as an illustration for such case. This clearly demonstrates how well the recall scores are enhanced when incorporating SSC information. For example, when plugged into SIFT it shows to increase enormously the recalls. Thus, in Tab. 5.1a, we read the scores of 0.05, 0.00, 0.00, and 0.01 for SIFT, while their correspondences obtained with SIFT-Based-SSC are 0.43, 0.39, 0.40, and 0.41, respectively.

These results confirm the performance gain, in terms of discriminative power, when adding the SSC information.

Furthermore, Figs. 5.6, 5.7, 5.8 and 5.9 show that for the high precision range above 0.9, *i.e.*, *1-precision* below 0.1, the three variants of the SSC descriptor, SIFT-based-SSC, SPIN-based-SSC and CC-based-SSC, still provide higher recalls while others produce too-low recall scores. These scores are practically zero, as example, in the case of Harris-Laplace regions with nearest-neighbor, as displayed in Fig. 5.6a.

The previous prominent performance of the SSC approach arise from two main reasons. The first is general and is related to the degree of scene similarities. Whereas the second is specific to the rotation. The SSC-based component is inherently rotation invariant.

TAB. 5.1: Evaluation of the recall scores under image rotation for the textured scene of *field*. The results are obtained for the descriptors matched using (a) *nearest-neighbor* and (b) *threshold-based matching* techniques. The recall scores are computed for the precision thresholds of (a) 0.90 and (b) 0.80.

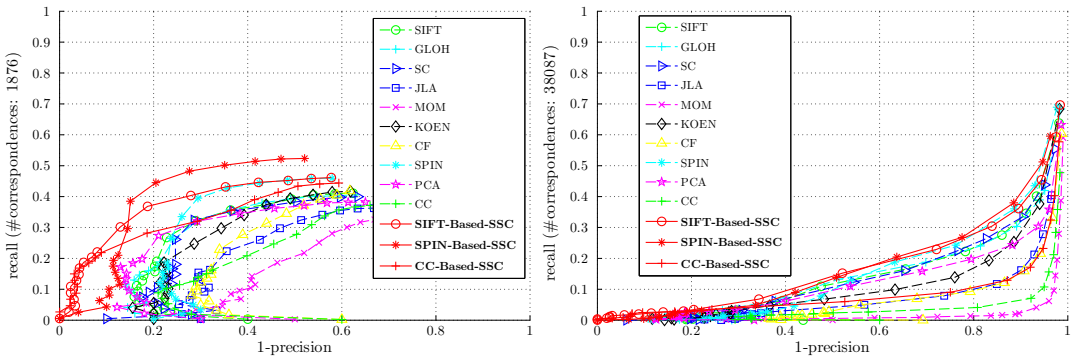
(a) *nearest-neighbor matching*

Descriptor	Harris-Laplace	Hessian-Laplace	Harris-Affine	Hessian-Affine
SIFT	0.05	0.00	0.00	0.01
GLOH	0.30	0.30	0.30	0.25
SC	0.29	0.30	0.30	0.31
SPIN	0.00	0.00	0.00	0.00
PCA	0.13	0.09	0.13	0.18
SIFT-Based-SSC	**0.43**	**0.39**	**0.40**	**0.41**
SPIN-Based-SSC	0.15	0.15	0.15	0.05

(b) *threshold-based matching*

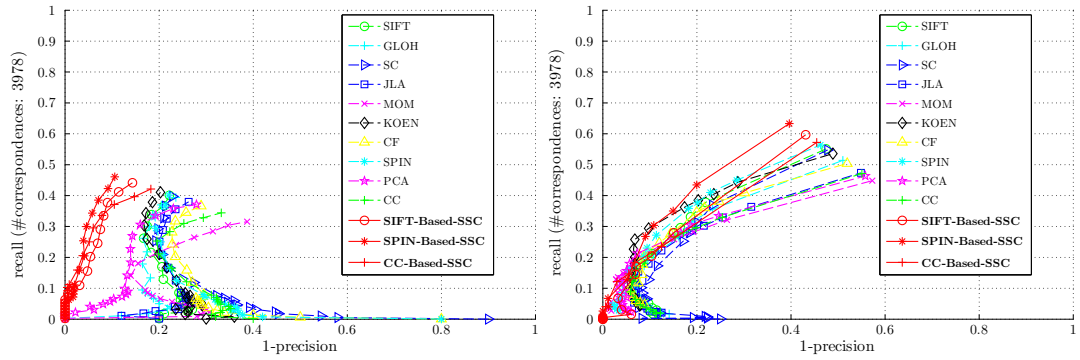
Descriptor	Harris-Laplace	Hessian-Laplace	Harris-Affine	Hessian-Affine
SIFT	0.00	0.00	0.00	0.00
GLOH	0.15	0.12	0.15	0.14
SC	0.15	0.12	0.15	0.15
SPIN	0.09	0.08	0.10	0.08
PCA	0.04	0.04	0.06	0.08
SIFT-Based-SSC	**0.20**	**0.16**	**0.20**	**0.20**
SPIN-Based-SSC	0.09	0.08	0.10	0.09

That is, the accumulated occurrences of each semantic feature inside each concentric shell are nearly unchanged under rotation. Hence, the resulting histograms are notably insensitive to arbitrary rotations.



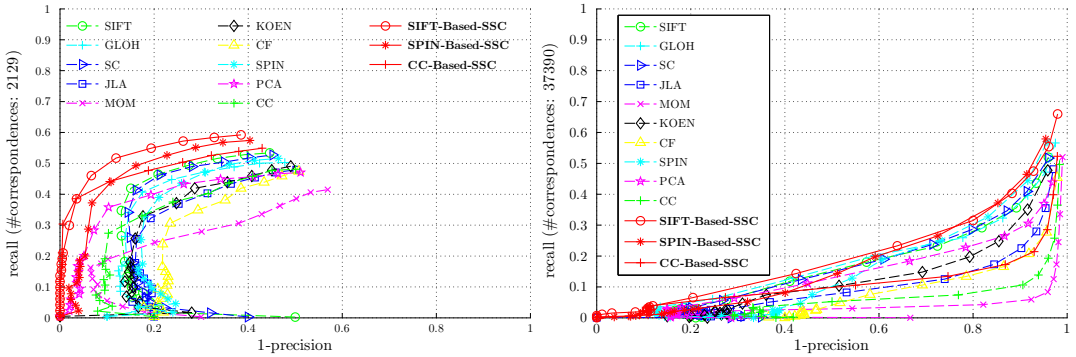
(a) (b)

FIG. 5.6: Evaluation results of the discriminative power under image rotation. The results are obtained for the structured scene, *chessboard* of Fig. 5.1a. The descriptors are computed for Harris-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1876 and (b) 38087 correspondences.



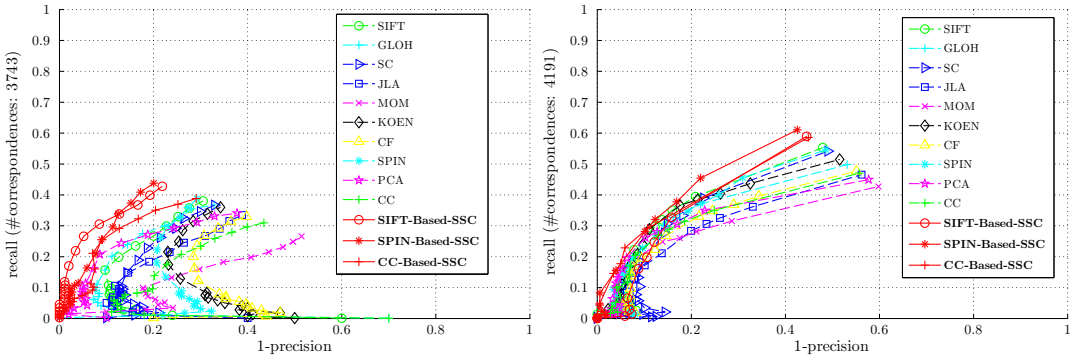
(a) (b)

FIG. 5.7: Evaluation results of the discriminative power under image rotation. The results are obtained for the structured scene, *chessboard* of Fig. 5.1a. The descriptors are computed for Hessian-Laplace regions and then matched using the (a) nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are computed with respect to 3978 correspondences.



(a) (b)

FIG. 5.8: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *chessboard* of Fig. 5.1a. The descriptors are computed for Harris-Affine regions and then matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2129 and (b) 37390 correspondences.



(a) (b)

FIG. 5.9: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *chessboard*, of Fig. 5.1a. The descriptors are computed for Hessian-Affine regions and then matched using (a) the nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are computed with respect to (a) 3743 and (b) 4191 correspondences.

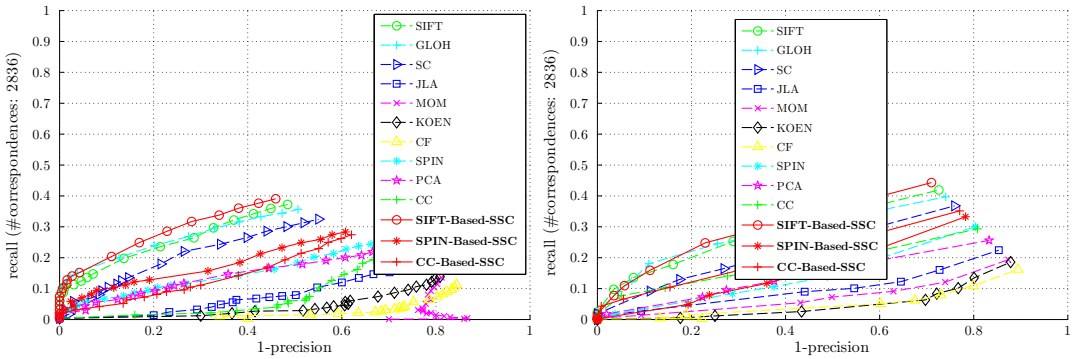


FIG. 5.10: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *streets* of Fig. 5.1b. The descriptors are computed for Harris-Laplace regions and matched using (a) nearest-neighbor and (b) distance-ration-based nearest-neighbor matching techniques. The recall scores are computed with respect to 2836 correspondences.

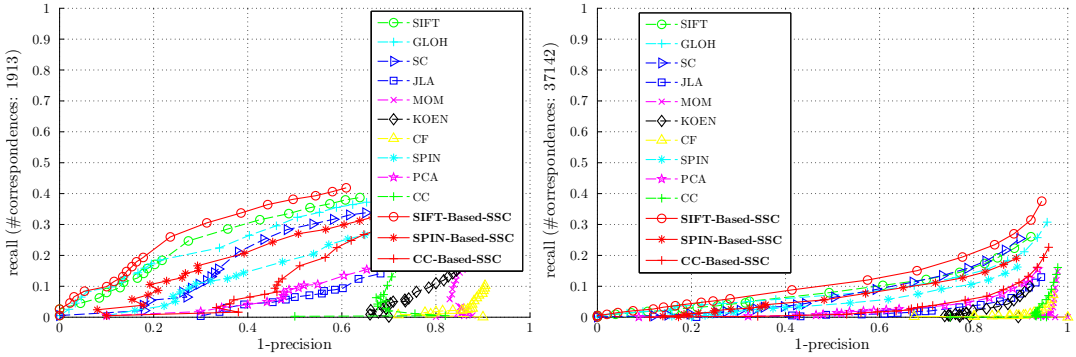


FIG. 5.11: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *streets* of Fig. 5.1b. The descriptors are computed for Hessian-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1913 and (b) 37142 correspondences.

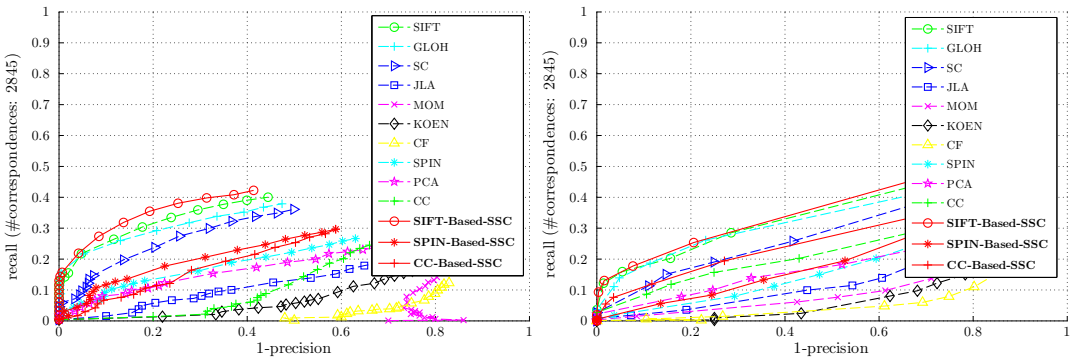


FIG. 5.12: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *streets* of Fig. 5.1b. The descriptors are computed for Harris-Affine regions and matched using (a) the nearest-neighbor and (b) distance-ration-based nearest-neighbor matching techniques. The recall scores are computed with respect to 2845 correspondences.

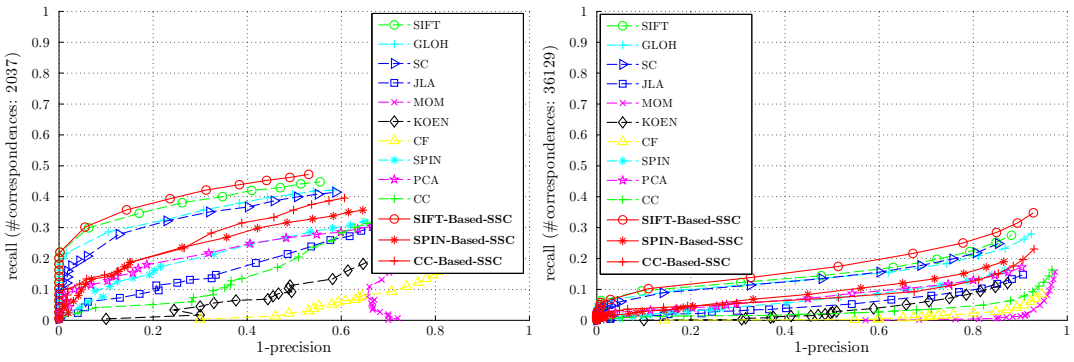
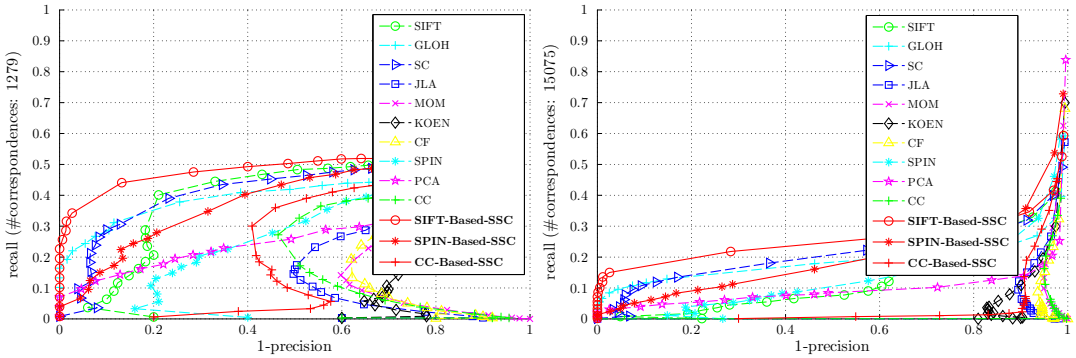
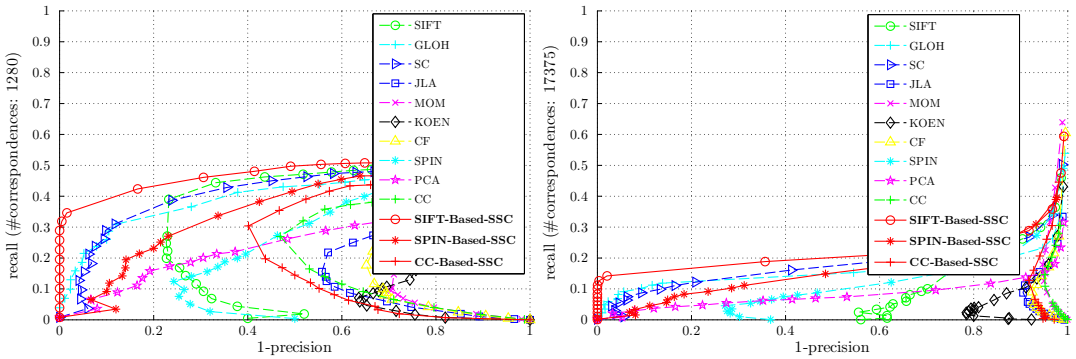


FIG. 5.13: Evaluation results of the discriminative power for image rotation. The results are obtained for the structured scene, *streets* of Fig. 5.1b. The descriptors are computed for Hessian-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2037 (b) 36129 correspondences.



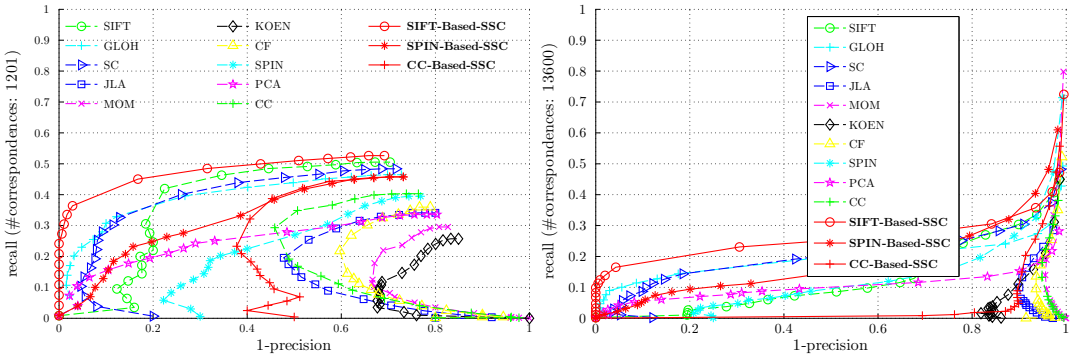
(a) (b)

FIG. 5.14: Evaluation results of the discriminative power for image rotation. The results are obtained for the textured scene, *field* of Fig. 5.1c. The descriptors are computed for Harris-Laplace regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1279 and (b) 15075 correspondences.



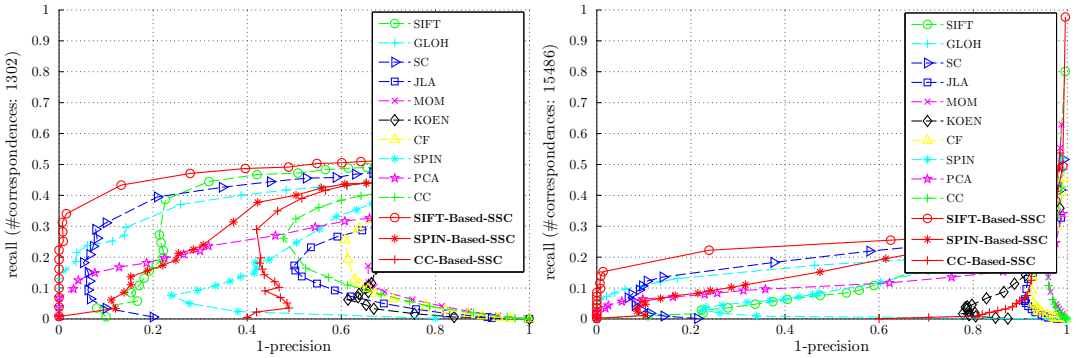
(a) (b)

FIG. 5.15: Evaluation results of the discriminative power against image rotation. The results are obtained for the structured scene, *field* of Fig. 5.1c. The descriptors are computed for Hessian-Laplace regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1280 and (b) 17375 correspondences.



(a) (b)

FIG. 5.16: Evaluation results of the discriminative power for image rotation. The results are obtained for the textured scene, *field* of Fig. 5.1c. The descriptors are computed for Harris-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1201 and (b) 13600 correspondences.



(a) (b)

FIG. 5.17: Evaluation results of the discriminative power for image rotation. The results are obtained for the textured scene, *field* of Fig. 5.1c. The descriptors are computed for Hessian-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed for (a) 1302 and (b) 15486 correspondences.

5.3.3.2 Invariance

In this experiment, we evaluate the invariance of each descriptor under image rotation. This is performed by measuring the constancy of the recall and precision scores for gradual increases in image rotation.

The evaluations are conducted on the scene of a *field*, which contains 6 images with different rotation angles ranged from 5° to 25°. The descriptors computed for `harris-affine` and `hessian-affine` regions, and matched using both `1-nearest-neighbor` and `threshold-based` matching techniques. The obtained results are displayed in Figs. 5.18 and 5.19.

In general, results show SIFT-Based-SSC recording the best constancy (robustness) under all experimental settings for both recall and precision. Besides, the invariance of SPIN and CC is considerably enhanced when adding the SSC information.

In detail, as expected, the recall and precision scores decrease when increasing image rotation, whereas it is less important for SIFT-Based-SSC than the other descriptors. Thus, we observe all SIFT-Based-SSC curves drop down slowly than those of the other descriptors.

As example, Fig. 5.18a shows while the recall score of SIFT-Based-SSC drops from 0.6 to 0.44, other descriptors like, SIFT, SC and GLOH, drop from 0.6 to 0.09, 0.0, and 0.3,respectively. A similar observation can be made with precision scores, as seen in Figs. 5.18c, 5.18d, 5.19c, and 5.19d. Therefore, the lack in discriminative power of SIFT is compensated in SIFT-Based-SSC which still provide a high discriminative power even for large rotation angles.

Finally, we notice that the descriptors ranking is approximately non preserved within evaluations. However, SIFT-Base-SSC always obtains the top spot followed by SC in the case of `nearest-neighbor` and by GLOH in the case of `threshold-based` matching. It is worth noting the better ranking of SPIN-Based-SSC compared to SIFT, as shown in Figs. 5.18b and 5.19b.

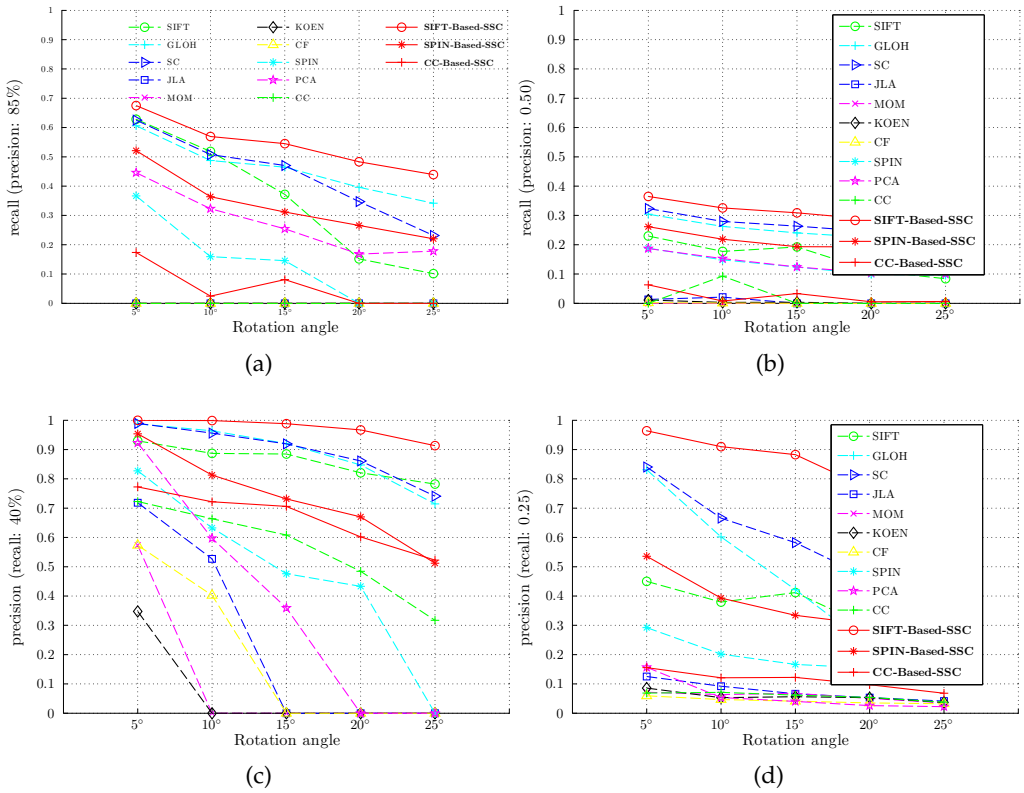


FIG. 5.18: Evaluation results of the invariance under image rotation. The results are obtained for the textured scene, *field* of Fig. 5.1c. The descriptors are computed for Harris-Affine regions and matched using the (a)(c) nearest-neighbor and (b)(d) threshold-based techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.85 and (b) 0.50. The precision scores are computed with respect to the recall thresholds of (c) 0.40 and (d) 0.25.

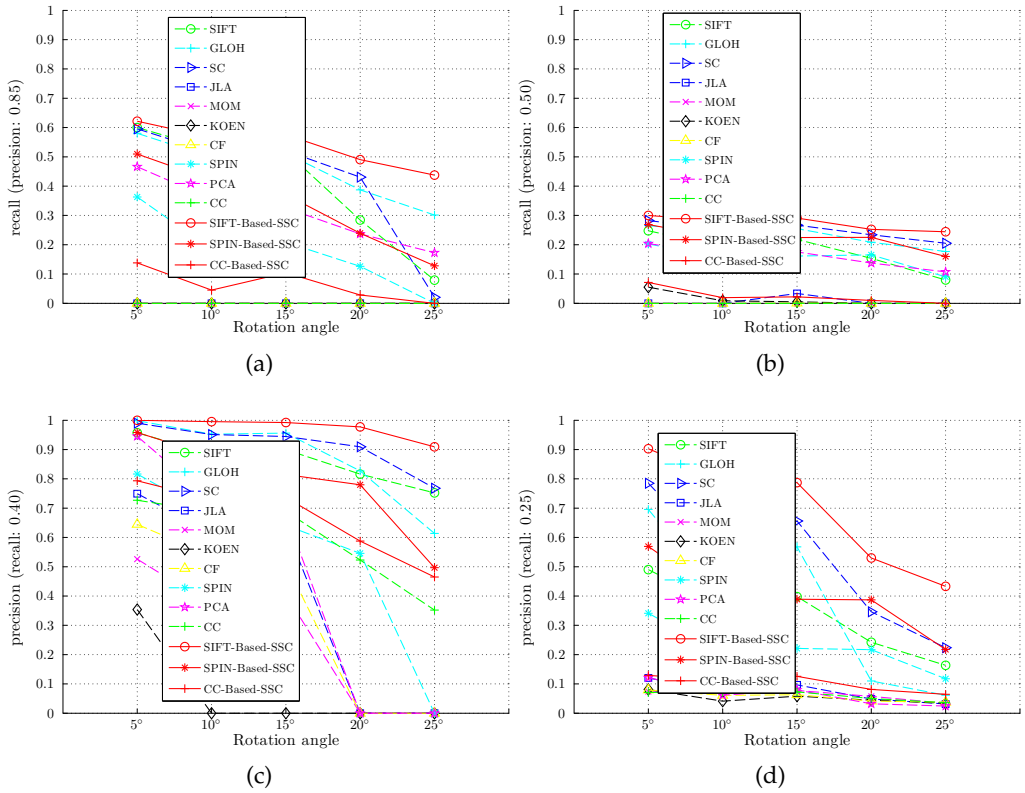


FIG. 5.19: Evaluation results of the invariance under image rotation. The results are obtained for the textured scene, *field* of Fig. 5.1c. The descriptors are computed for Hessian-Affine regions and matched using the (a)(c) nearest-neighbor and (b)(d) threshold-based techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.85 and (b) 0.50. The precision scores are computed with respect to the recall thresholds of (c) 0.40 and (d) 0.25.

5.3.4 Scale Change

For scale change, we evaluate the descriptor performances by examining both discriminative power and invariance in two scenes. The first is the `giraffe` of Fig. 5.1i, which contains 9 images obtained approximately with scale factors of range 1.2-3.5. The second is `grass` of Fig. 5.1j, which is composed of 3 images with scale factors of 1, 0.75 and 0.65.

Regarding discriminative power, the descriptors are evaluated for different support regions, as well as for `nearest-neighbor` and `threshold-based` matching techniques.

Since scale change is more challenging than image rotation, the discriminative power is evaluated under both low and high scale factors. Whereas for the invariance, the descriptors are computed for `hessian-laplace` and `hessian-affine` detectors and then matched with `nearest-neighbor` and `threshold-based` matching techniques.

An example of `nearest-neighbor` matching using SIFT-Based-SSC for the `giraffe` and `grass` scenes are shown in Figs. 5.20a and 5.20b.

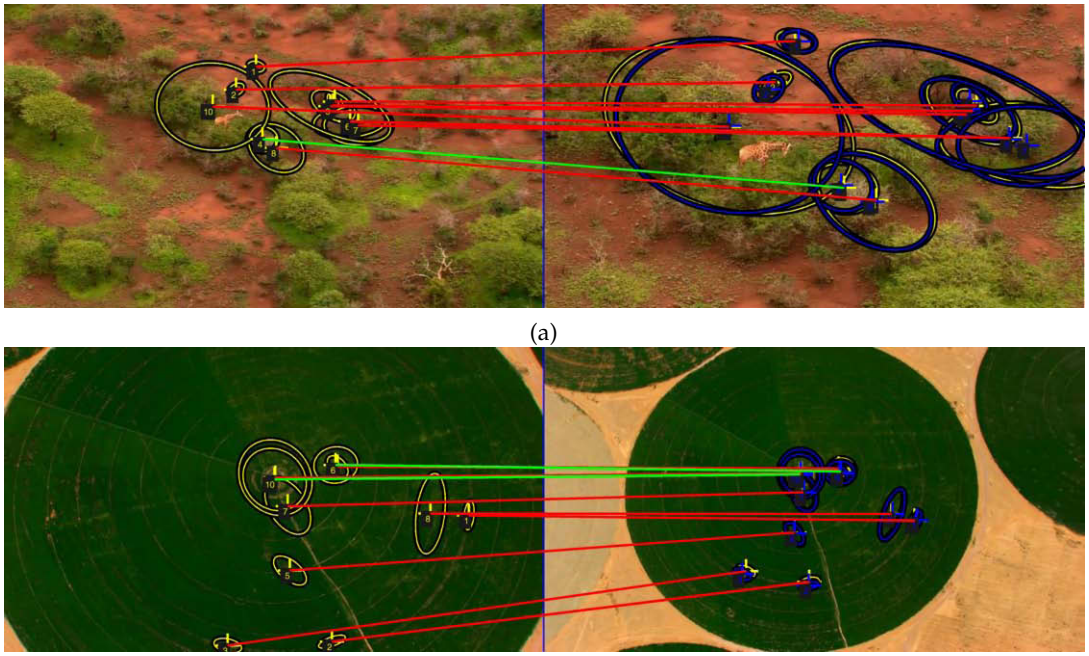
5.3.4.1 Discriminative power

We investigate the discriminative power under scale change for the descriptors computed on different support regions and matching strategies as well. The experiments are performed on the scenes of `giraffe` and `grass`.

For the scene of `giraffe`, the discriminative power is evaluated for images of low (≈ 1.3) and high (≈ 2) scale factors. The observations are focused more on the left side of ROC graphs. The obtained results are reported in Figs. 5.21-5.28.

At first glance, SIFT-Based-SSC seems to perform mostly better than the other descriptors. The gap between SIFT-Baesd-SSC and the others become extremely large for images with high scale factors as displayed in Figs. 5.21a, 5.22a, 5.23a, and 5.24a. It is also observed that CC-Based-SSC performs much better than CC for all experimental settings, where it turns out to outperform some competitive approaches like PCA-SIFT, for example.

In more details, although well known to perform better under scale change in the textured scene like, `giraffe`, SC appears to be fully outperformed by SIFT-Based-SSC for all region detectors and matching strategies as well. The results, such as shown in Figs. 5.21a, 5.21b, 5.22a and 5.22b, clarify the large difference between SIFT-Based-SSC



(a)

(b)

FIG. 5.20: An example of nearest-neighbor matching using SIFT-Based-SSC computed for Harris-Affine regions. The results are obtained for the (a) textured and (b) structured scenes. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines. For the purpose of clarity, only a reduced number of regions are shown.

ranked first and the SC ranked second. It is important to observe the SIFT ranking before and after incorporating the SSC information. It jumped from 3, or 4, to the first spot.

It is also interesting to watch attentively the impact of plugging the SSC information into CC. The resulting highly improved CC-Based-SSC becomes a workable descriptor while obtaining a competitive discriminative power.

It is not the case for the CC descriptors computed in images with high scale factors as shown in Figs. 5.21a, 5.21b. These appear completely unworkable, since they are unable to produce any correct match even for low 1-precision scores.

For instance, Tab. 5.2, lists the 1-precision/recall thresholds for which the CC descriptors seem to be not operational. In order words they are not able to produce correct matches which have precision scores below reported thresholds.

TAB. 5.2: The 1-precision/recall thresholds for which the CC descriptor achieves its limitations. That is, without producing any correct matches. The results are obtained for the image of *giraffe* with scale factor of 2.

Matching method	Harris-Laplace	Hessian-Laplace	Harris-Affine	Hessian-Affine
nearest-neighbor matching	0.40/0.22	0.70/0.22	0.70/0.27	0.65/0.20
Threshold-based matching	0.80/0.03	0.92/0.01	0.97/0.05	0.90/0.04

We note that the above tabled scores are reported for 1-precision and recall values of 0.60 and 0.40 respectively.

According to the scene type, the obtained results on the `grass` scene, show GLOH coming in the second rank behind SIFT-Based-SSC, while SIFT obtains the third rank followed by SC.

We observe for `hessian-laplace`, *i.e.*, Fig. 5.22, SIFT-Based-SSC and GLOH perform almost equally while SIFT is completely outperformed. It has recall scores of zero for 1-precision scores below 0.1. Which means the considerable impact of SSC information in improving the SIFT discriminative power.

We notice that SSC-based descriptors obtain better discriminative powers on `giraffe` than on `grass`. In addition to the higher self similarities of `giraffe`, the other reason is because of spatial distribution of similar motifs inside `giraffe`'s scene. The random repartition of region similarities without any uniform structures reinforces disambiguation in SSC component which in turn enhances the discriminative power.

The contribution of the SSC information into our descriptor is more relevant in `girraffe` than in `grass`. This is because the clustering in the first is more stable and accurate. Thus, the semantic features represented by the clusters are obtained for the first scene

from 9 images across large scale factors (*i.e.*, range of $1.2-3.5$) while for the second are obtained from 3 images within small scale factors (*i.e.*, range of $0.65-0.75$).

This is related to the fact that the role of the semantic information (which is generated from visual features) is keeping tracks of informations that can always appear in an image at different scales. Hence, more the images and large scale factors, the more the semantic information.

In conclusion, under scale change, the discriminative power of SIFT is greatly enhanced when adding SSC information. It is also more than expected to obtain an improved performance of CC-Based-SSC when SSC is plugged into CC. The SSC information plays a primordial role in improving the descriptor discriminative power under image scaling.

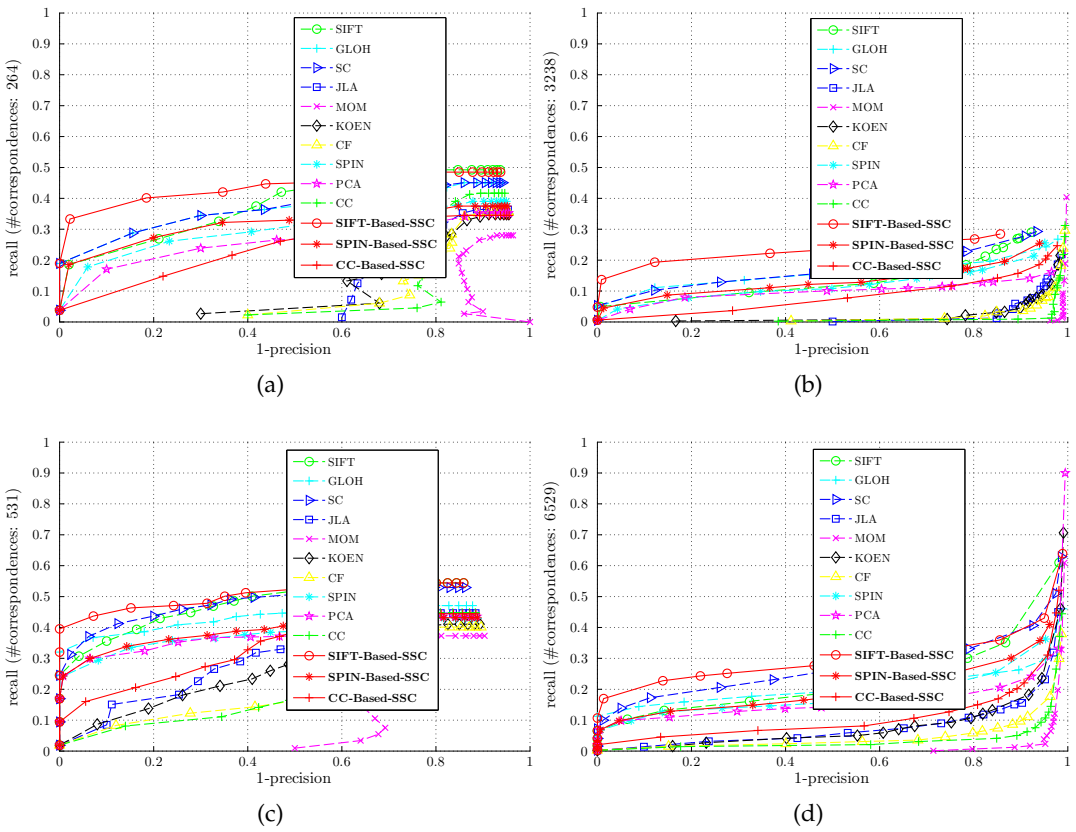


FIG. 5.21: Evaluation results of the discriminative power under scale change. The results are shown for images with (a)(b) high and (c)(d) low scale factors of the scene, *giraffe* (Fig. 5.1i). The descriptors are computed for Harris-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 264, (b) 3238, (c) 531, and (d) 6529 correspondences.

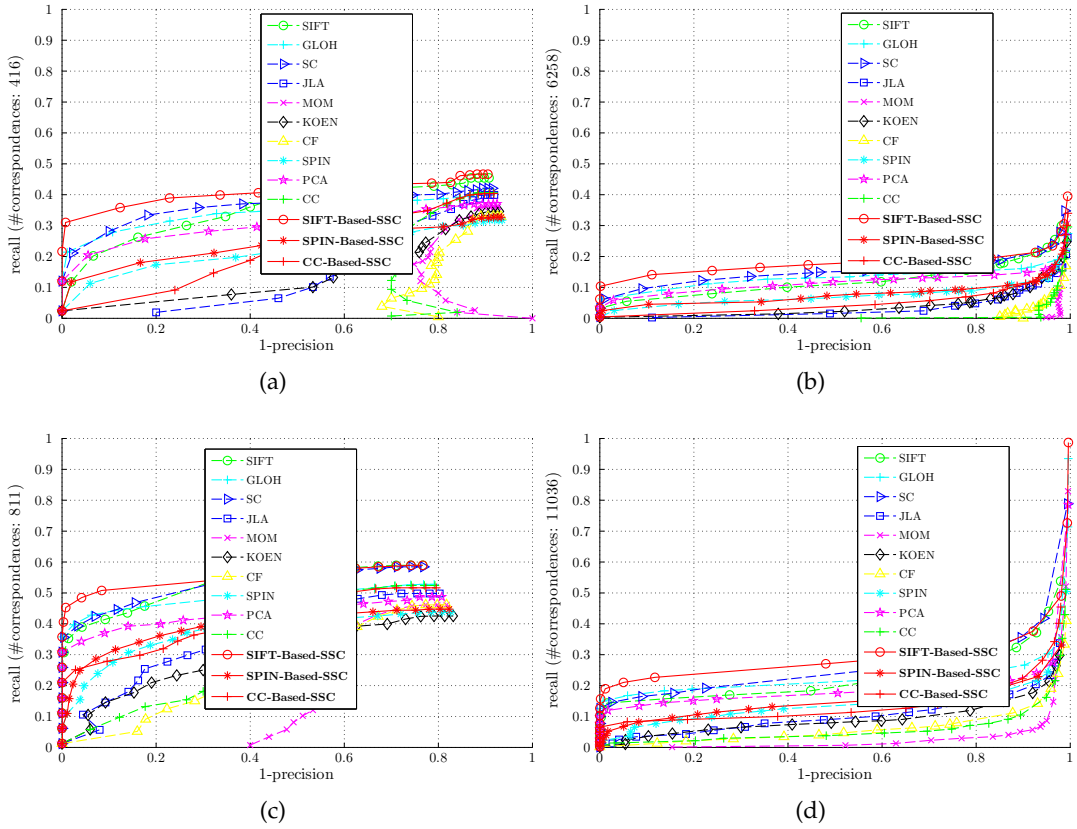


FIG. 5.22: Evaluation results of the discriminative power under scale change. The results are shown for image with (a)(b) high and (c)(d) low scale factors of the structured scene of *giraffe* (Fig. 5.1i). The descriptors are computed for Hessian-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 416, (b) 6258, (c) 811, and (d) 11036 correspondences.

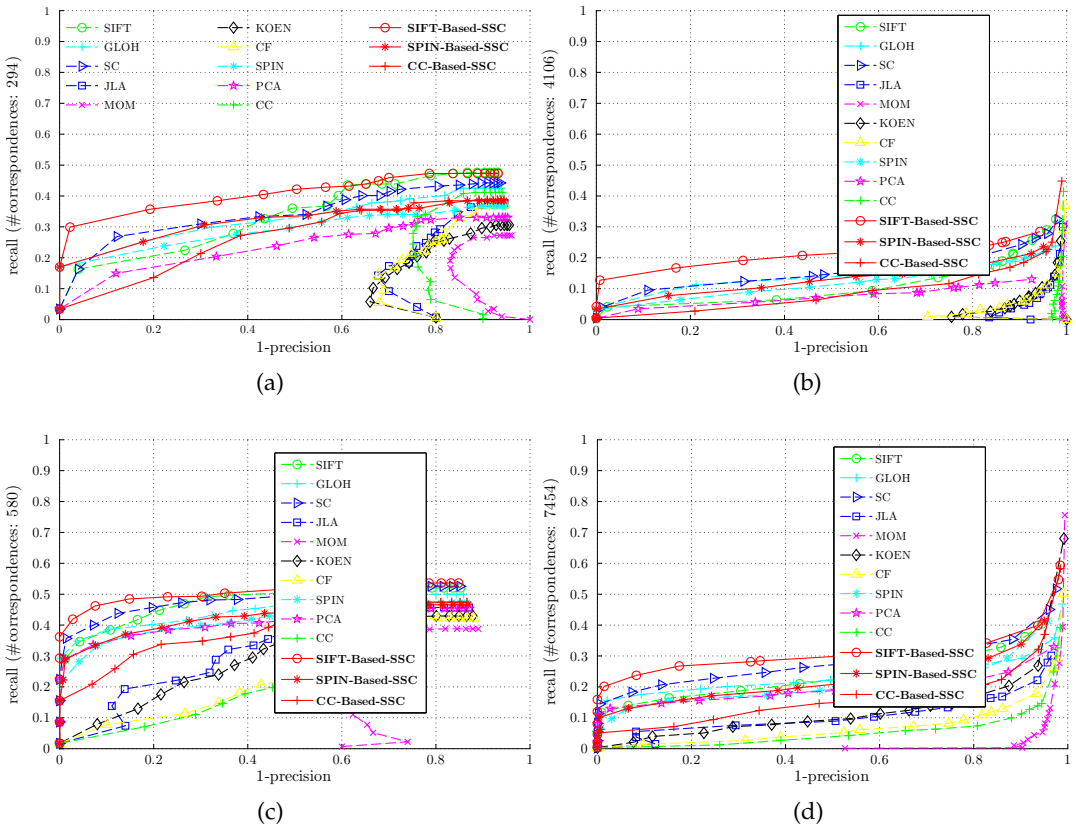


FIG. 5.23: Evaluation results of the discriminative power under scale change. The results are shown for images with (a)(b) high and (c)(d) low scale factors of the textured scene, *giraffe* (Fig. 5.1i). The descriptors are computed for Harris-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 294, (b) 4106, (c) 580, and (d) 7454 correspondences.

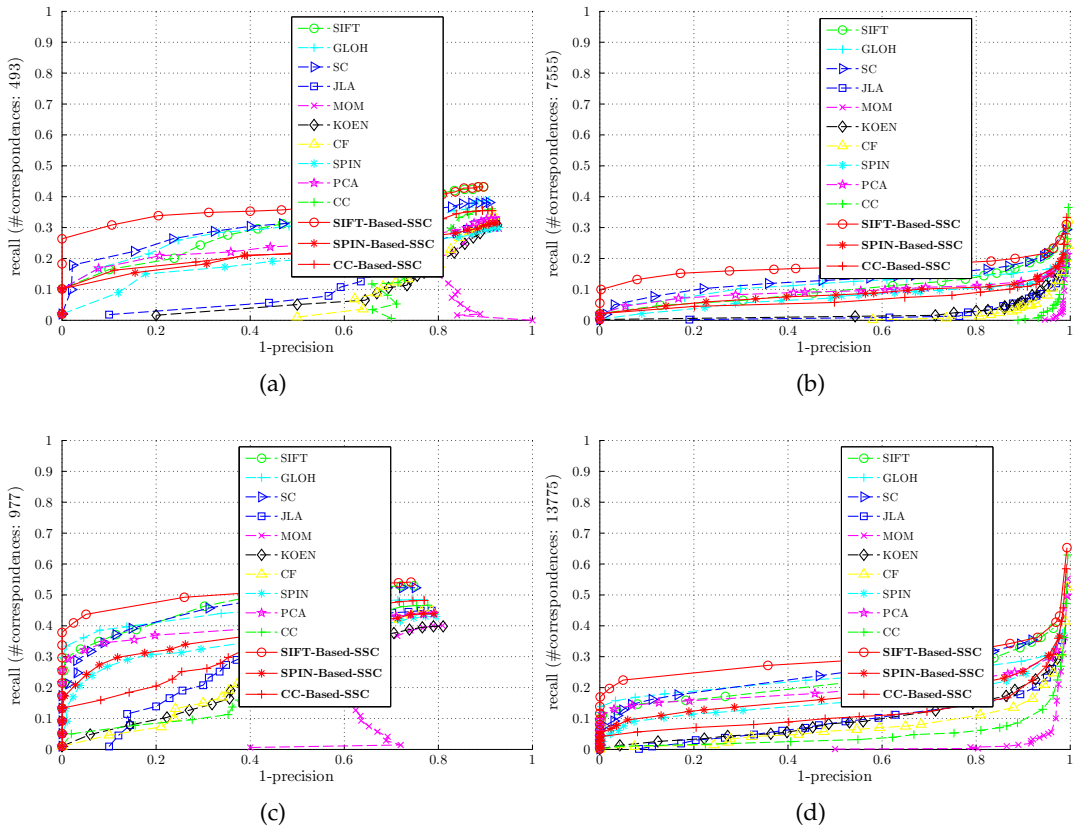
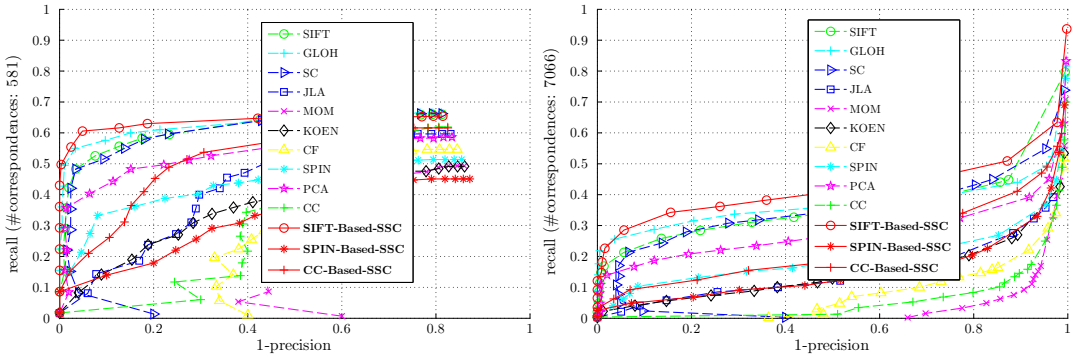
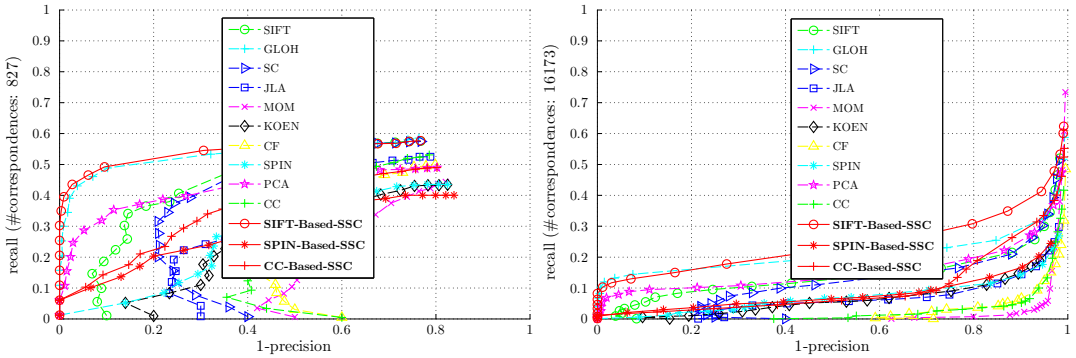


FIG. 5.24: Evaluation results of the discriminative power under scale change. The results are shown for images with (a)(b) high and (c)(d) low scale factors of the scene, *giraffe* of Fig. 5.1i. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 493, (b) 7555, and (c) 977, and (d) 13775 correspondences.



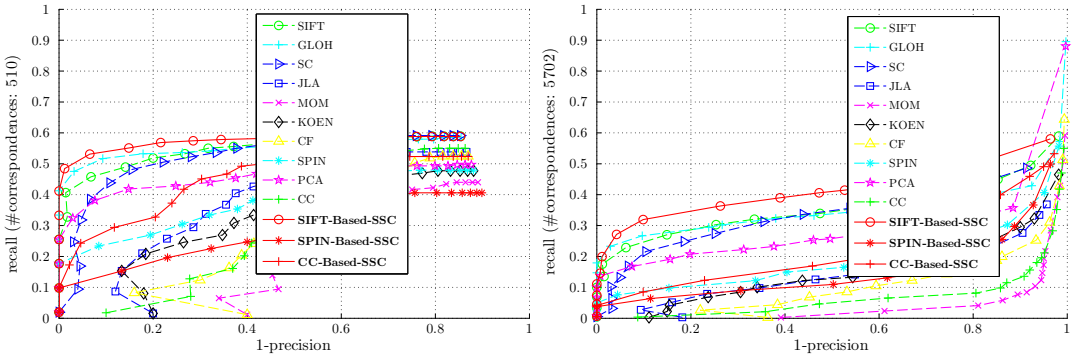
(a) (b)

FIG. 5.25: Evaluation results of the discriminative power under scale change. The results are obtained for the scene, *grass*, of Fig. 5.1j. The descriptors are computed for Harris-Laplace regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 581 and (b) 7066 correspondences.



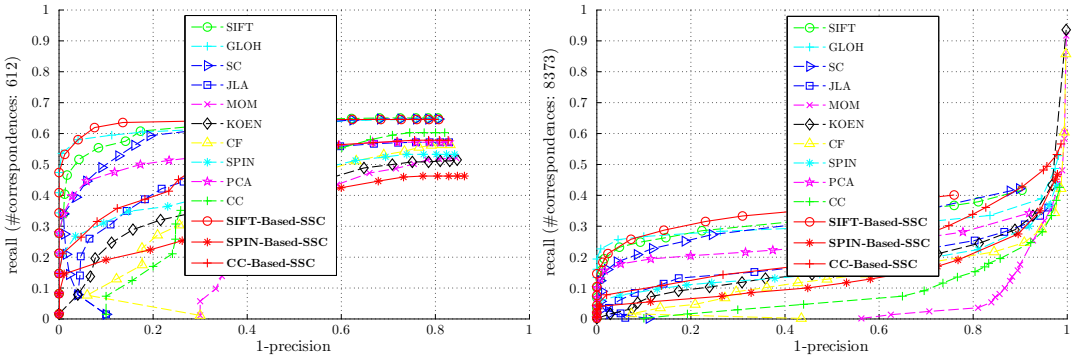
(a) (b)

FIG. 5.26: Evaluation results of the discriminative power under scale change. The results are obtained for the scene, *grass* of Fig. 5.1j. The descriptors are computed for Hessian-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 827 and (b) 16173 correspondences.



(a) (b)

FIG. 5.27: Evaluation results of the discriminative power under scale change. The results are obtained for the scene, *grass* of Fig. 5.1j. The descriptors are computed for Harris-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 510 and (b) 5702 correspondences.



(a) (b)

FIG. 5.28: Evaluation results of the discriminative power under scale change. The results are obtained for the scene of *grass* of Fig. 5.1j. The descriptors are computed for Hessian-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 612 and (b) 8373 correspondences.

5.3.4.2 Invariance

In these experiments, the descriptor invariance for scale changes is evaluated through inspecting the degradation of both recall and precision scores under progressive increases in the scale factor.

The evaluations are conducted on the scene of `giraffe`, which contains 8 images with different scale factors varying from 1.2 to 3.5. Since image scaling is more attractive than rotation, the invariance is measured for descriptors computed for two different region detectors, `hessian-laplace` and `hessian-affine`. For each region detector, the recall and precision scores are computed for both `1-nearest-neighbor` and `threshold-based` matching methods. The results are given through Figs. 5.29 and 5.30.

The reported figures clearly highlight the effectiveness of semantic-context information on the descriptor invariance. Thus, SIFT invariance is highly enhanced when semantic-context information is added.

This can be checked by observing the ranking of SIFT and SIFT-Based-SSC curves within Figs. 5.29 and 5.30. It seems SIFT-Based-SSC curves drop down more slowly than those of other descriptors. This is shown for SIFT-Based-SSC which always occupies the first position for both recall and precision over all scale factors. Besides, SC, GLOH and SIFT ranked second, third , and fourth respectively.

Usually considered well suited for the textured scene, it appears that the SC is completley won by SIFT-Based-SSC. This situation is more obvious for the precision invariance as highlighted in Figs. 5.29c, 5.29d, 5.30c and 5.30d.

Tab. 5.3 summarizes the degradations in the recall and precision scores from the low to high scale factors, *i.e.*, from 1.2 to 3.5. These scores are obtained for the four best descriptors computed for `hessian-laplace` and matched with the `nearest-neighbor` method.

TAB. 5.3: Degradations in recall and precision scores under scale change obtained for the scene of *giraffe*. The degradations are computed from low to high scale factors (*i.e.*, from 1.2 to 3.5) of recall (a) and (b) precision scores.

(a) Recall

Descriptor	low scale	high scale	*degradation*
SIFT	**0.62**	0.00	0.62
SC	**0.62**	0.09	0.53
GLOH	0.58	0.00	0.58
SIFT-Based-SSC	**0.62**	**0.20**	**0.42**

(b) Precision

Descriptor	low scale	high scale	*degradation*
SIFT	**1.00**	0.30	0.70
SC	**1.00**	0.34	0.66
GLOH	**1.00**	0.15	0.85
SIFT-Based-SSC	**1.00**	**0.70**	**0.30**

This table shows that SIFT-Based-SSC recorded the smallest degradation for both recall

and precision. For the recall, we count a degradation of 0.42 with SIFT-Based-SSC while we obtain 0.62, 0.53, and 0.58 with SIFT, SC, and GLOH, respectively. For precision, the gap is huge in the sense that we register only a degradation of 0.30 for SIFT-Based-SSC and the best of the remaining has 0.66 with SC, as example.

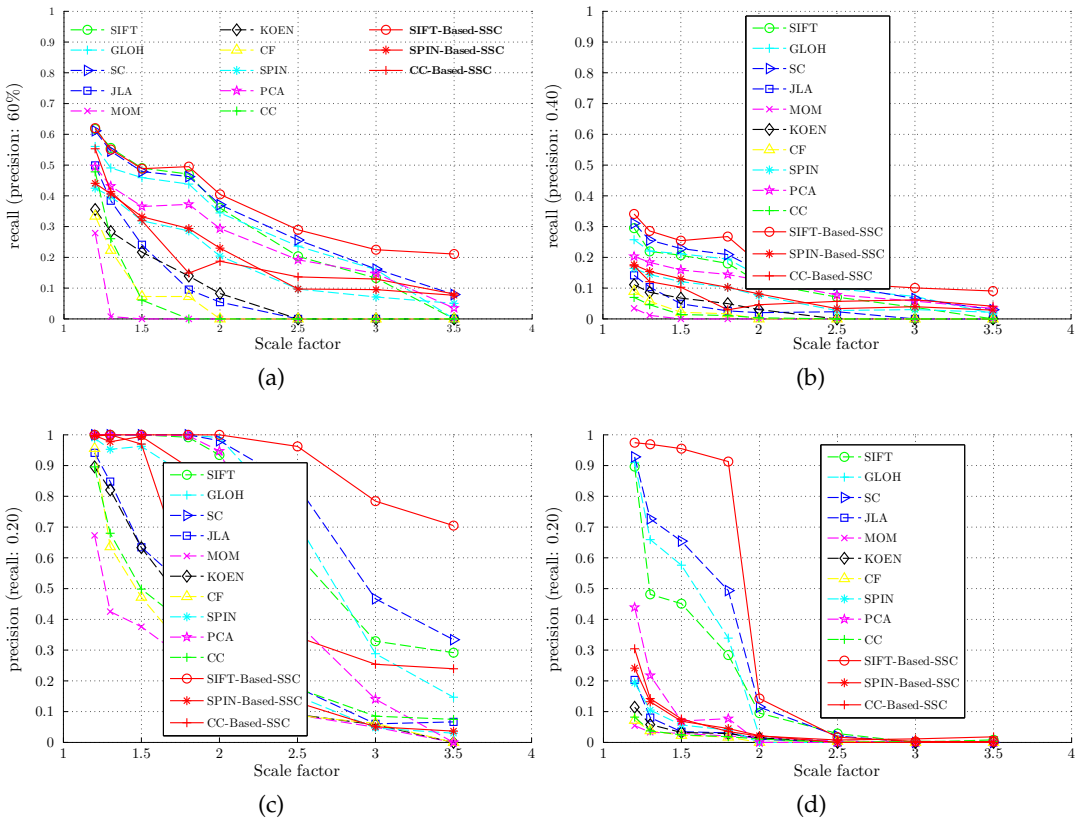


FIG. 5.29: Invariance evaluation under scale changes (range of $1.2-3.5$). The results are obtained for the structured scene, *giraffe* of Fig. 5.1i. The descriptors are computed for Hessian-Laplace regions and matched with (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.60 and (b) 0.40. The precision scores are computed with respect to the recall threshold of 0.20.

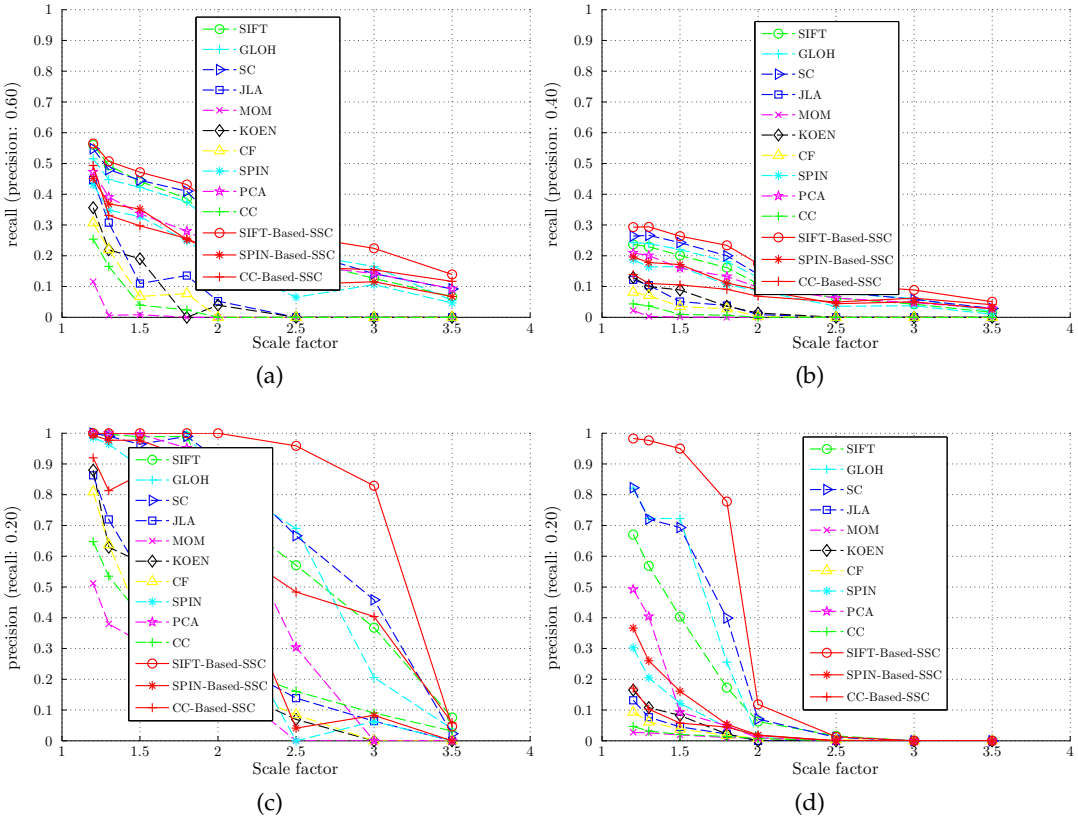


FIG. 5.30: Invariance evaluation under scale changes (range of $1.2-3.5$). The results are obtained for the structured scene of *giraffe* of Fig. 5.1i. The descriptors are computed for Hessian-Affine regions and matched with (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.60 and (b) 0.40. The precision scores are computed for the recall thresholds of 0.20.

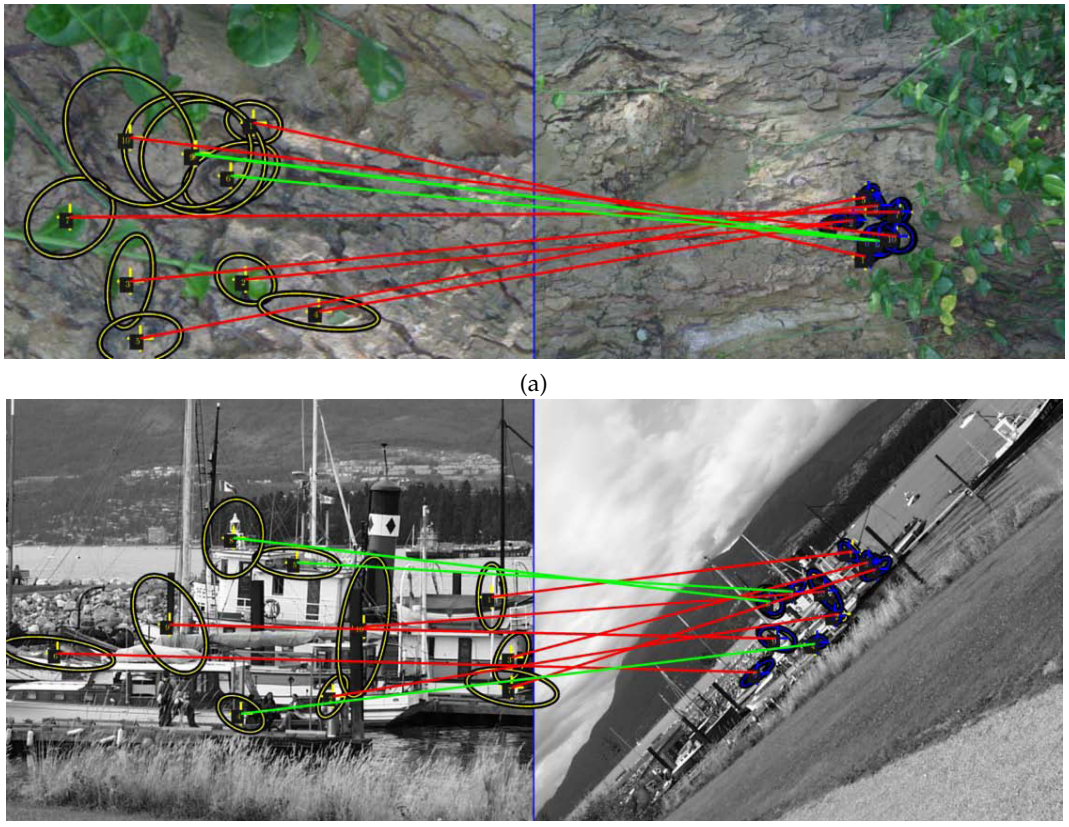
5.3.5 Rotation-enlargement transformation

In this experiment, the performance of the SSC approach is evaluated on scenes selected from the standard data set of *Mikolajczyk* described in Section 5.2.2. These scenes reflect the particular image deformation of rotation-enlargement transformation, which combines rotation and scale change.

For this purpose, we consider two different scenes. The first, `bark` of Fig. 5.2b, is a textured scene composed of 6 images. The second, `boat` of Fig. 5.2a, is a structured scene with the same number of images.

These scenes are used to assess the descriptor performance for different region detectors and matching strategies. In this context, we focus on the discriminative power performance only.

Before going into detail, we show through Fig. 5.31 an example of `nearest-neighbor` matching obtained for SIFT-Based-SSC computed on `harris-affine` regions.



(a)

(b)

FIG. 5.31: An example of nearest-neighbor matching using SIFT-Based-SSC computed on Harris-Affine regions. The results are obtained for the scene of (a) *bark* and (b) *boat*. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines. For the purpose of clarity, only a reduced number of regions are shown.

5.3.5.1 Discriminative power

Similar to scale changes, the discriminative power of the descriptor are compared for images representing small and large deformations. For the scene of `bark`, we use 3rd and 6th images, whereas for the `boat`, we consider the 3rd and 7th.

The descriptors of these images are computed for the four region detectors and matched with `nearest-neighbor` and `threshold-based` matching methods. The obtained results are given in Figs. 5.32-5.39.

At first look, we figure out that the best discriminative power is obtained by SIFT-Based-SSC descriptor. Although it performs differently within two scenes, SIFT-Based-SSC still ranked first in all situations. As expected, we obtain the best performance with the textured scene as shown in Figs. 5.34c and 5.34d.

Furthermore, we can observe how well the discriminative power of SIFT is enhanced for the structured scene when integrating the SSC component, *e.g.*, Figs. 5.36a, 5.36b, 5.38a, and 5.38b. This is more than expected since the scene of boat does not exhibit high region similarities.

The results demonstrate the usefulness of the SSC approach even for the structured ordinary scenes without multiple similar motifs.

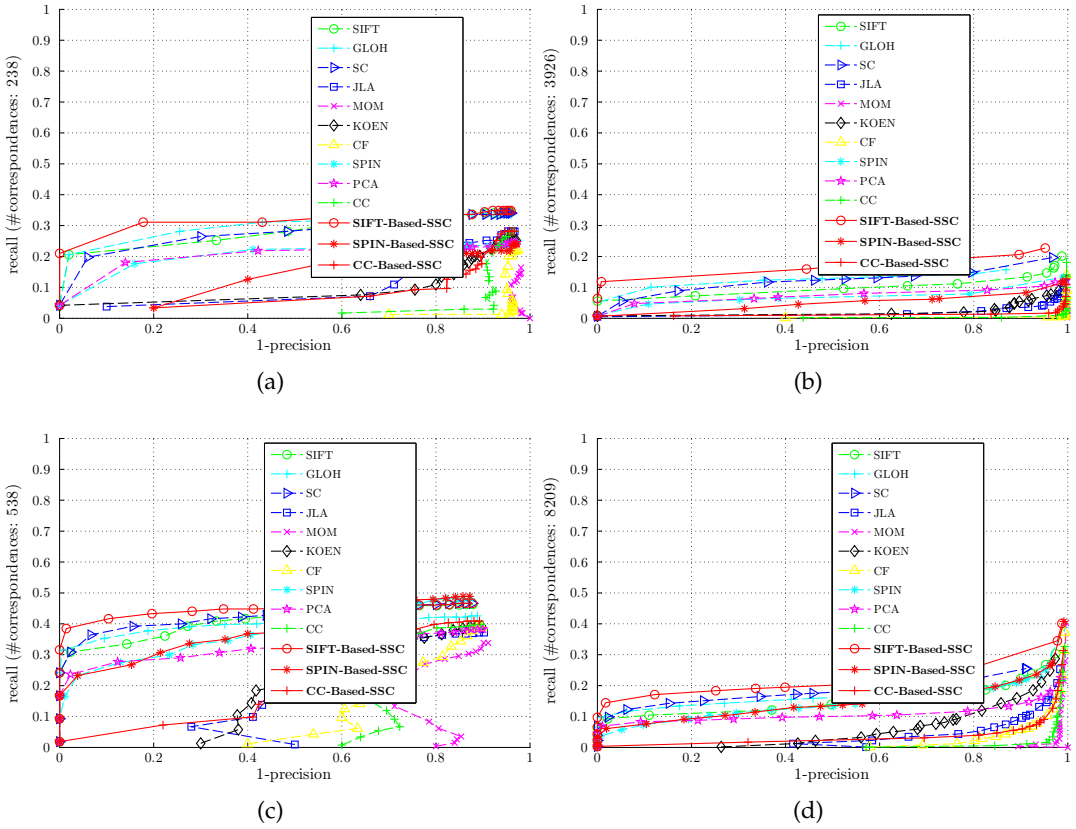


FIG. 5.32: Evaluation results of the discriminative power under rotation-enlargement. The results are shown for images with (a)(b) large and (c)(d) small rotation-enlargement of the scene, *bark* of Fig. 5.2b. The descriptors are computed for Harris-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 238, (b) 3926, (c) 538, and (d) 8209 correspondences.

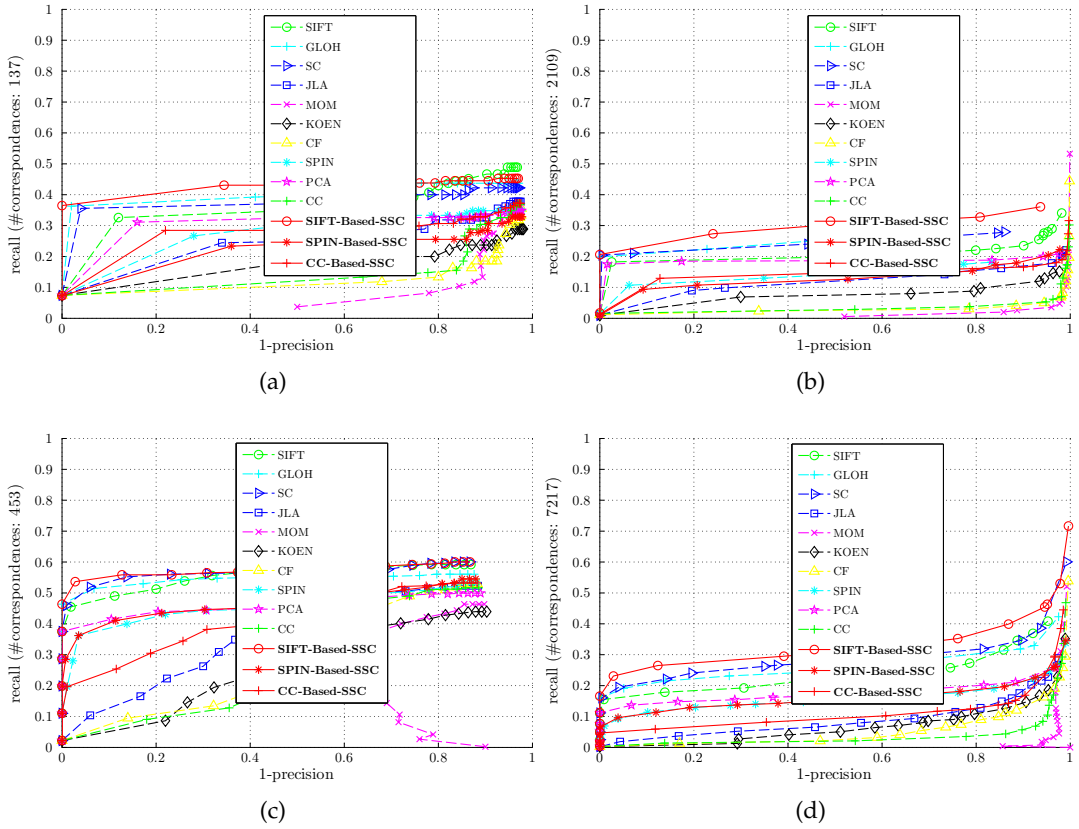


FIG. 5.33: Evaluation results of the discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the structured scene of *bark* of Fig. 5.2b. The descriptors are computed for Hessian-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 137, (b) 2109, (c) 453, and (d) 7217 correspondences.

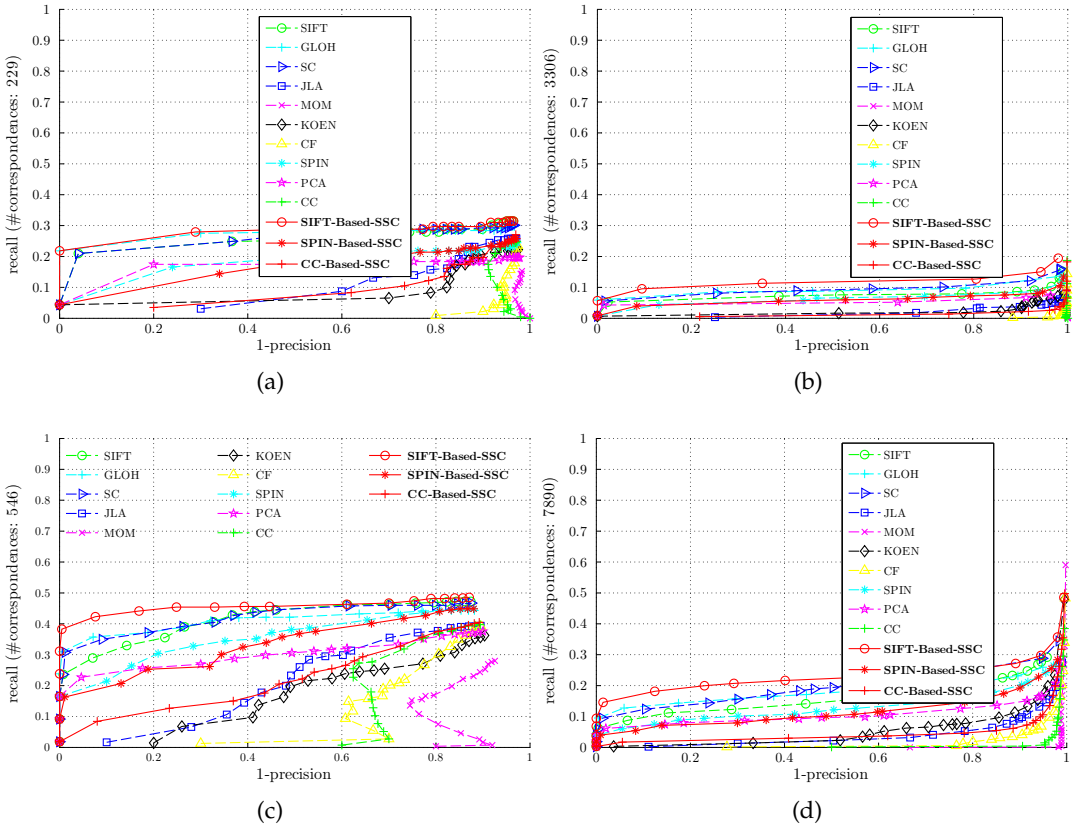


FIG. 5.34: Evaluation results of the discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the textured scene of *bark* of Fig. 5.2b. The descriptors are computed for Harris-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 229, (b) 3306, (c) 546, and (d) 7890 correspondences.

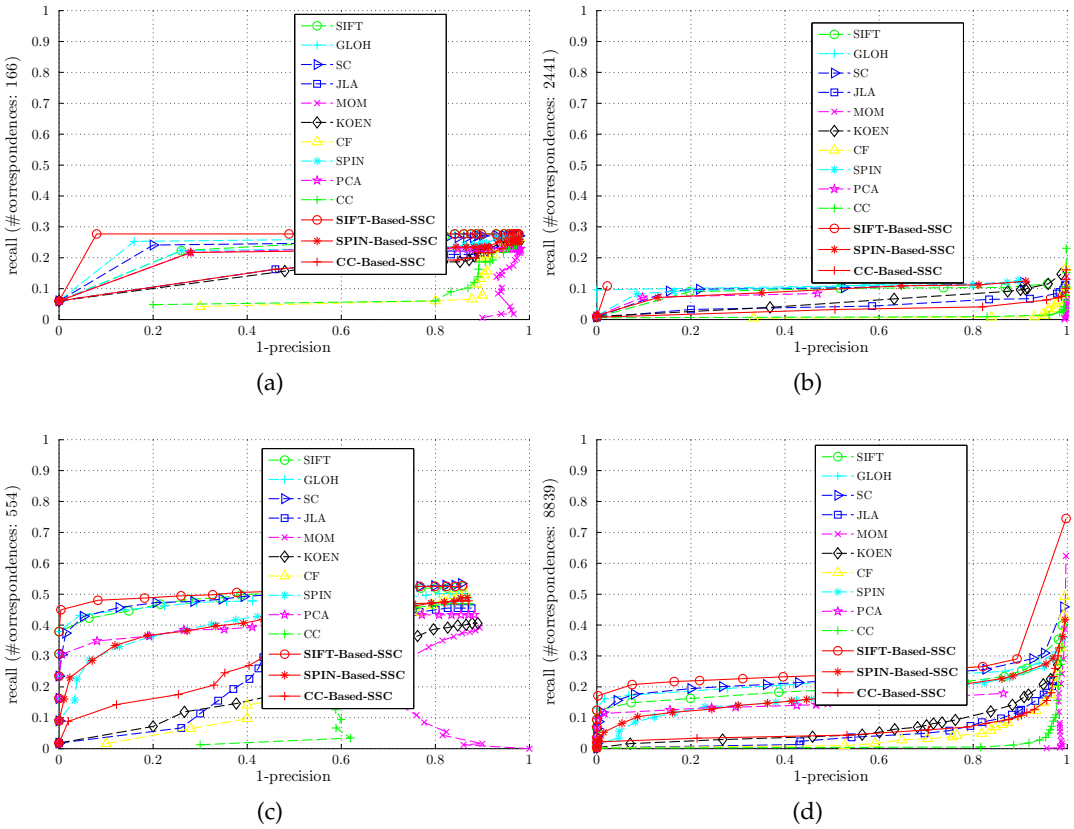


FIG. 5.35: Evaluation results of the discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the scene of Fig. 5.2b of the textured scene, *bark* of Fig. 5.2b. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 166, (b) 2441, (c) 554, and (d) 8839 correspondences.

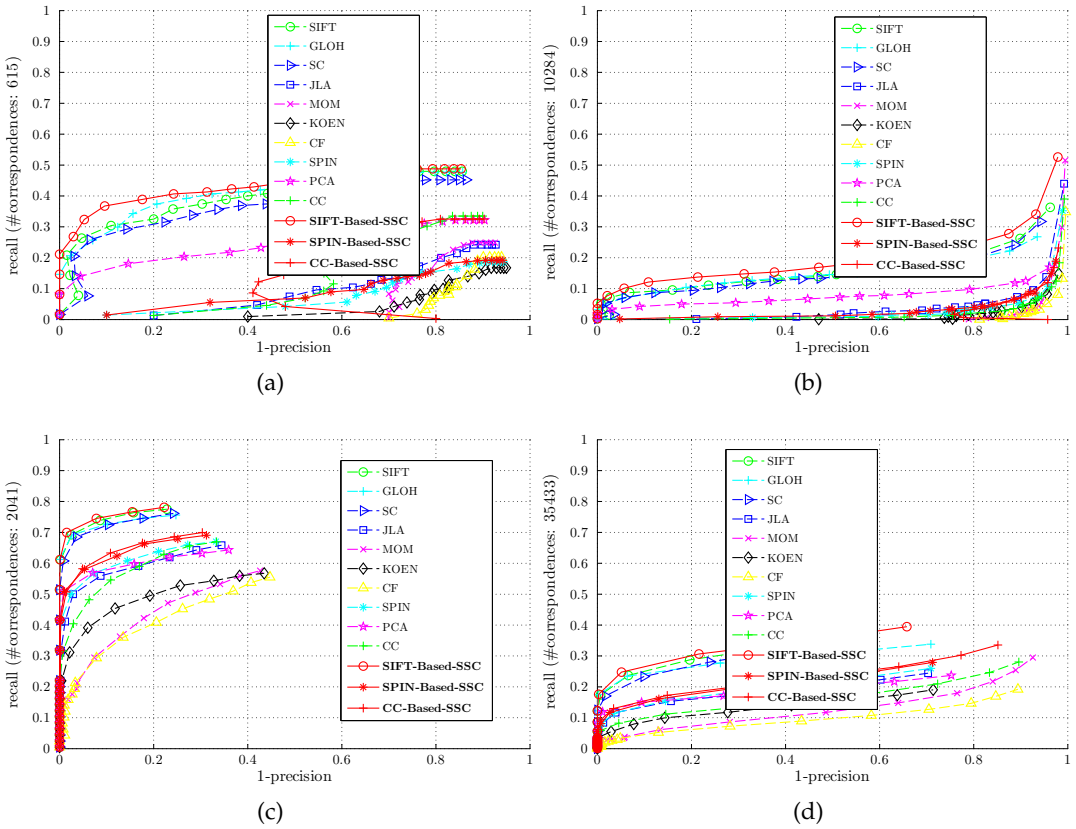


FIG. 5.36: Evaluation results of the discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) small rotation-enlargement of the structured scene, *boat*, shown in Fig. 5.2a. The descriptors are computed for Harris-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 615, (b) 10284, (c) 2041, (d) 35433 correspondences.

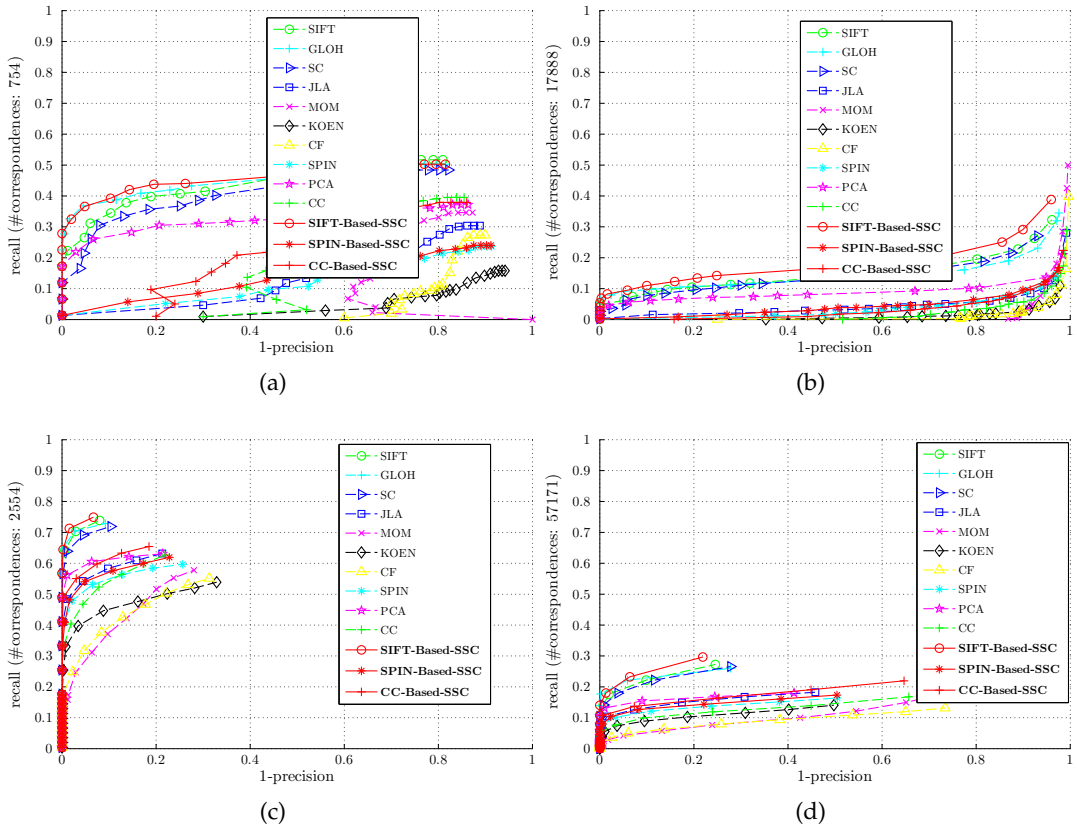


FIG. 5.37: Evaluation results of discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the structured scene, *boat*, shown in Fig. 5.2a. The descriptors are computed for Hessian-Laplace regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 754 , (b) 17888, (c) 2554, and 57171 (d) correspondences.

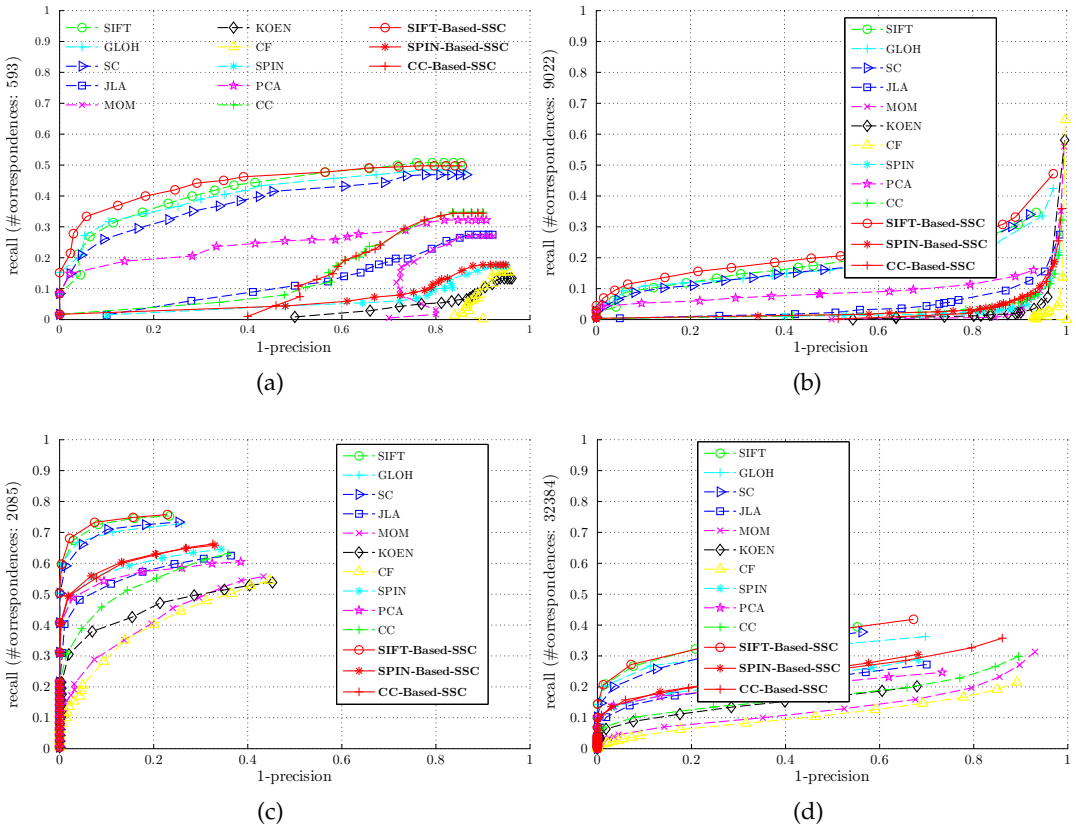


FIG. 5.38: Evaluation results of discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the structured scene, *boat*, shown in Fig. 5.2a. The descriptors are computed for Harris-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 593, (b) 9022, (c) 2085, and (d) 32384 correspondences.

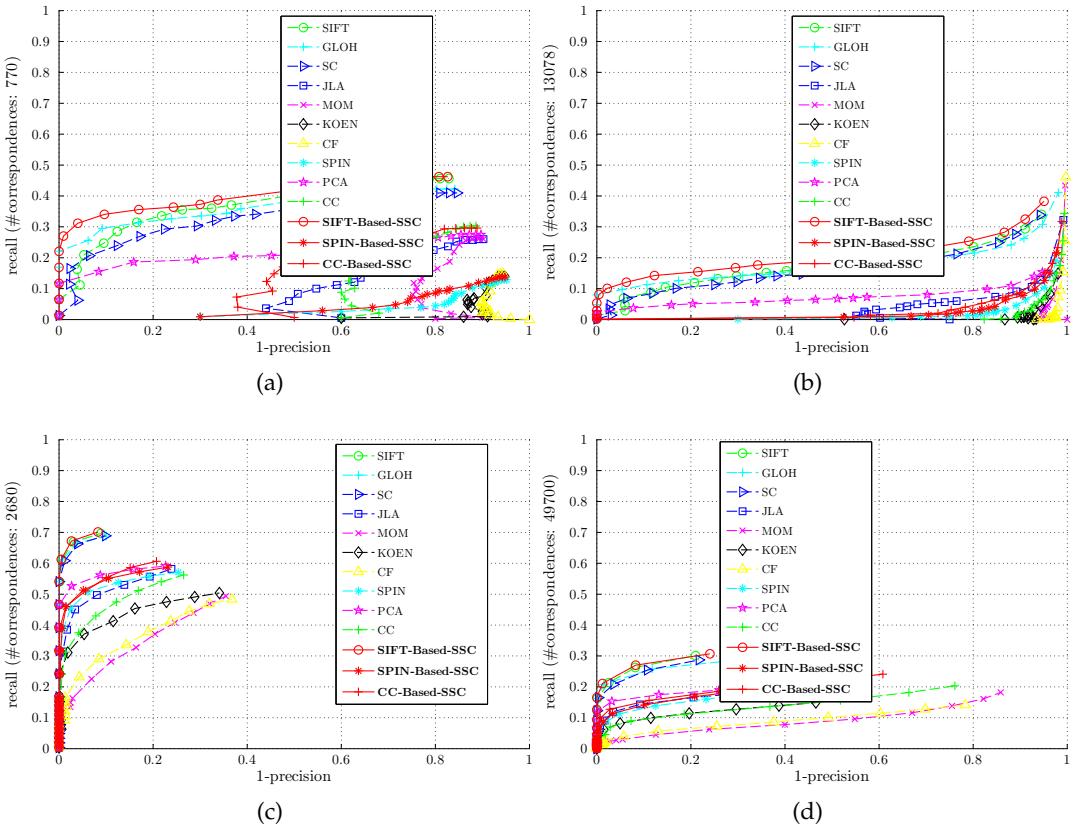


FIG. 5.39: Evaluation results of discriminative power under rotation-enlargement. The results are shown for images with (a)(b) high and (c)(d) low rotation-enlargement of the structured scene, *boat*, shown in Fig. 5.2a. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 770, (b) 13078, (c) 2680, and (d) 49700 correspondences.

5.3.6 Viewpoint Change

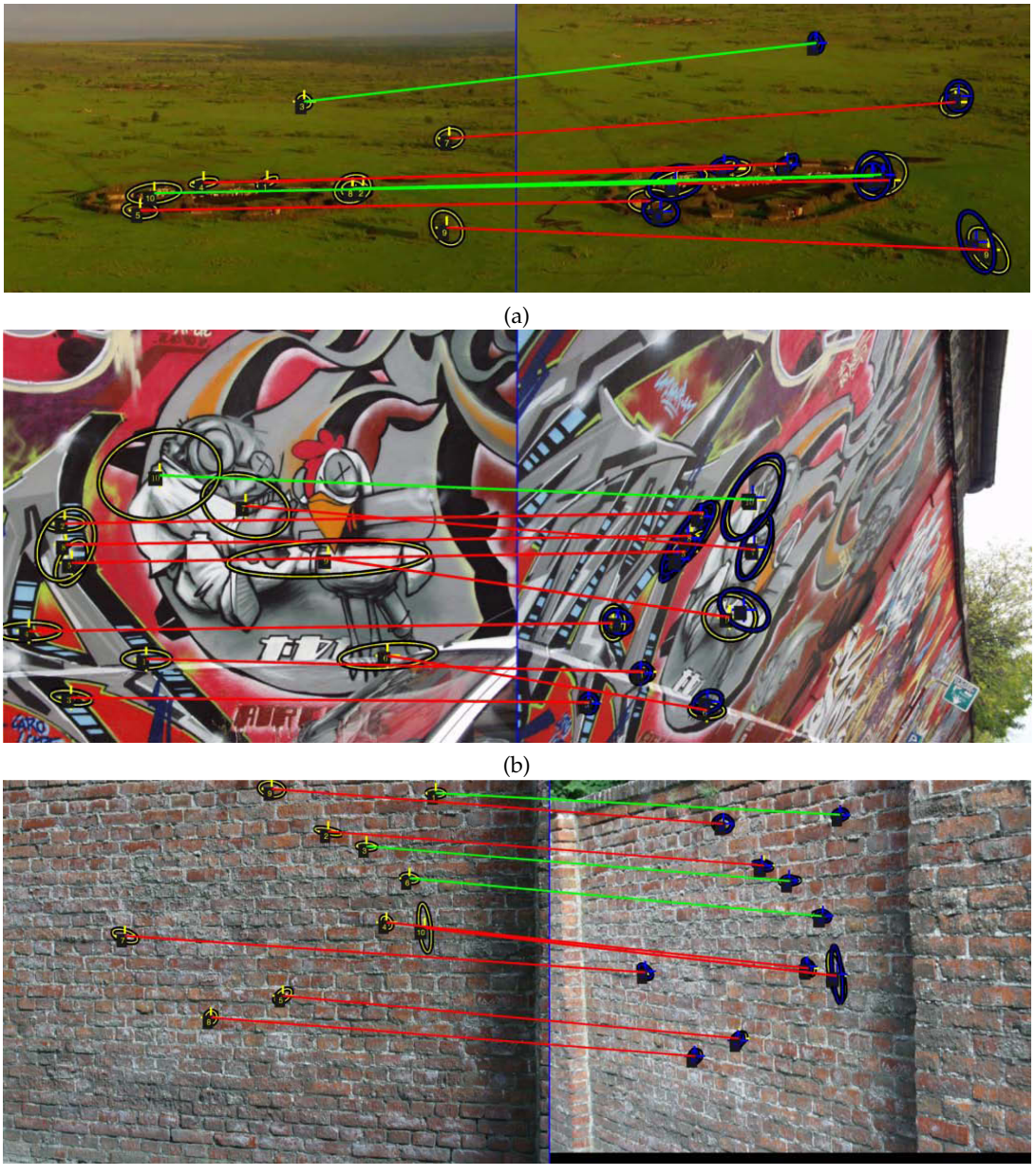
The viewpoint change is the most challenging deformation among different geometric image transformations. Therefore, we will pay much closer attention on their performance evaluation. In this context, we use three scenes expressing out-of-plane image rotations, *i.e.*, viewpoint changes.

The first scene, `zeriba` of Fig. 5.1k, is composed of 7 images of progressive increases in viewpoint angles from 10° to 70°. The second, `graffiti` of Fig. 5.2c, is a structured scene obtained from the standard data set of *Mikolajczyk*. This contains 5 images with viewpoint angles ranged in $20° - 60°$. The third, `wall` of Fig. 5.2d, represents a textured scene of 6 images selected also from *Mikolajczyk*'s dataset.

An example of `nearest-neighbor` matching for the descriptors computed for these scenes are shown in Fig. 5.40. These are obtained with SIFT-Based-SSC computed for `harris-affine` support regions.

The evaluation of the descriptors are performed according to the discriminative power and invariance. For the discriminative power, we use all scenes by considering the descriptors computed on the affine support regions and different matching methods. Whereas for the invariance, we use two scenes and the descriptors are also computed on the affine support regions and different matching algorithms as well.

In order to inspect the impact of the overlap error on the performance coherency, we start first by diagnosing the descriptor performances for different thresholds of overlap errors.



(a)

(b)

(c)

FIG. 5.40: An examples of nearest-neighbor matching using SIFT-Based-SSC computed for Harris-Affine regions. The results are obtained for the (a) textured scene, *zeriba* of Fig. 5.1k, (b) structured scene, *graffiti* of Fig. 5.2c, and (c) textured scene, *wall* of fig5.2d. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines. For the purpose of clarity, only a reduced number of regions are displayed.

5.3.6.1 Region overlap error

The main objective of this experiment is evaluating how the region overlap errors affect the descriptor performances. Thus, we evaluate changes in the recall and precision scores under different amounts of overlap errors ranged in $10\% - 60\%$.

We use two scenes of different types. These are the textured of `zeriba` and structured of `graffiti`. The descriptors are computed for `harris-affine` and `hessian-affine` affine detectors and matched using `1-nearest-neighbor` and `threshold-based` matching techniques.

In order to keep the coherence within all experiments and unlike [Mikolajczyk 05a] [4], we consider the score obtained for an overlap error, for example 20%, as that corresponding to an overlap error smaller than 20%. We therefore obtain an increased number of correspondences while increasing overlap error as highlighted in Tab. 5.4. The evalauation results related to the overlap errors are displayed within Figs. 5.41-5.44.

TAB. 5.4: Number of correspondences for different overlap errors obtained for the scene of *zeriba*. The scores are computed for (a) Harris-Affine and (b) Hessian-Affine region detectors.

(a) Harris-Affine

Matching method	10%	20%	30%	40%	50%	60%
nearest-neighbor	8	73	330	695	1059	1273
Threshold-based	21	241	1573	5343	14623	31636

(b) Hessian-Affine

Matching method	10%	20%	30%	40%	50%	60%
nearest-neighbor	5	136	561	1140	1611	1897
Threshold-based	12	485	3343	12086	34484	79676

Before going into details, we observe that the obtained curve shapes differ from one scene to another as well as from one matching strategy to another.

Basically, we wait for the recall (or precision) score to increase while overlap error is still increased below a particular threshold from which it starts decreasing. This is because, the higher the overlap error the higher number of correspondences (*i.e.*, in reference to Tab. 5.4) and thus the higher probability to recognize more potential correct matches as correct.

This is also related to the errors resulting from region detectors (imprecision of regions)

[4] We think this is the more appropriate than considering that computed for an overlap error larger than 10% and smaller than 20% as in [Mikolajczyk 05a].

which can influence the number of correct matches, for example, a potential correct match which cannot be identified as correct because its region overlap error is larger than the used threshold although they are in correspondence.

The value of overlap error from which the scores start decreasing corresponds to that from which the probability of matching more descriptors begins to decrease. This value is higher for the `nearest-neighbor`, *e.g.*, Figs. 5.41a and 5.41c, than for the `threshold-based` method, *e.g.*, Figs. 5.41b and 5.41d. This is because, the first is mostly correct and always selects only the best match below threshold and rejects the remaining although for large overlap error.

However, the second selects many matches and many of them are false, we therefore obtain lesser correct matches (and higher false matches) when the overlap error starts increasing. The result is that the value of overlap error from which the recall and precision scores decrease is high with the `nearest-neighbor` and low with the `threshold-based` matching method.

We observe that for the textured scenes, *i.e.*, Figs. 5.41 and 5.42, the scores obtained with the `nearest-neighbor` are always increasing, even for large overlap errors. This is due to the fact that, this type of scene contains a large number of similar motifs resulting in large number of similar regions which are spatially close to each other, and thus the probability of finding more correct matches becomes higher especially for the `nearest-neighbor` which is mostly correct.

The obtained figures display how well the performance of descriptors is improved when the SSC information is used. This can be checked, for example, by observing the recall curves of CC descriptor in Fig. 5.41a before and after adding the SSC component.

This figure shows that for an overlap error of 40%, we obtain recall scores of zero with CC and ≈ 0.1 with CC-Based-SSC. The gap is more apparent for an overlap error of 60%, in which we count a recall of 0.1 for CC and 0.35 for CC-Based-SSC. Similar observations can be seen for the precision. As example, Fig. 5.42b demonstrates that for an overlap error of 30%, while CC-Based-SSC produces a precision score of 0.2, the CC gives ≈ 0.

The most impact of the SSC information on the descriptor performance under overlap error is obtained for the SIFT descriptor. This is illustrated in Figs. 5.41-5.44, in which the SIFT-Based-SSC conserves its first rank within all evaluations.

Moreover, by comparing the curves of Figs. 5.41 and 5.42 to those of Figs. 5.43 and 5.44, we figure out that SIFT-Based-SSC performs better in the textured scene.

In conclusion, the overlap error does not affect the descriptors ranking and performance is gained when SSC descriptors are used especially in textured scene.

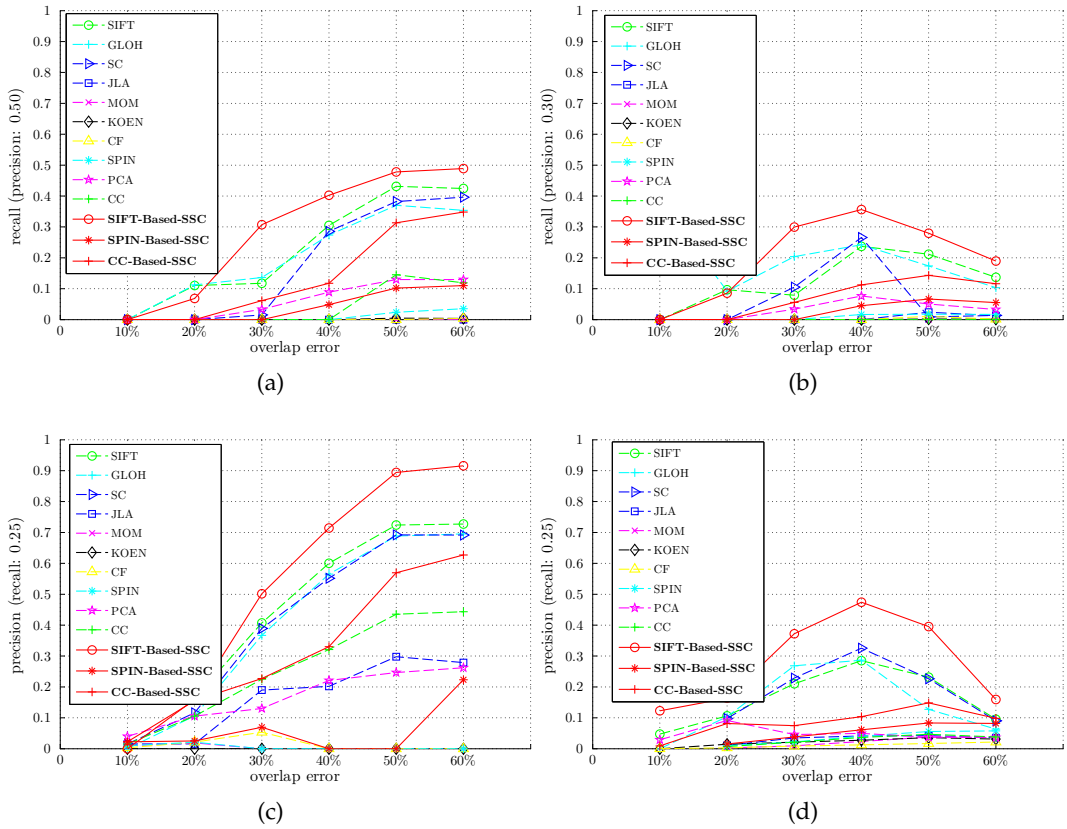


FIG. 5.41: Performance evaluation under viewpoint change for different overlap errors. The results are obtained for the textured scene, *zeriba* of Fig. 5.1k. The descriptors are computed for Harris-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed for the precision thresholds of (a) 0.50 and (b) 0.30. The precision scores are computed for the recall thresholds of (c) 0.25 and (d) 0.25.

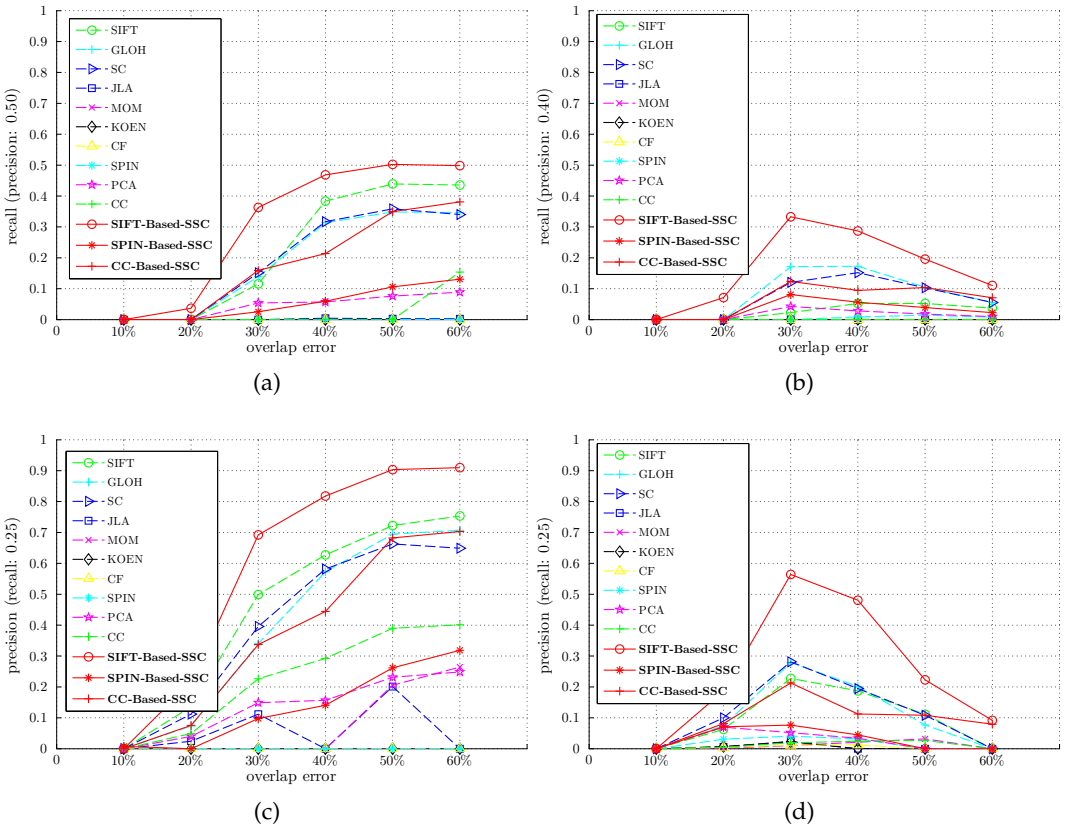
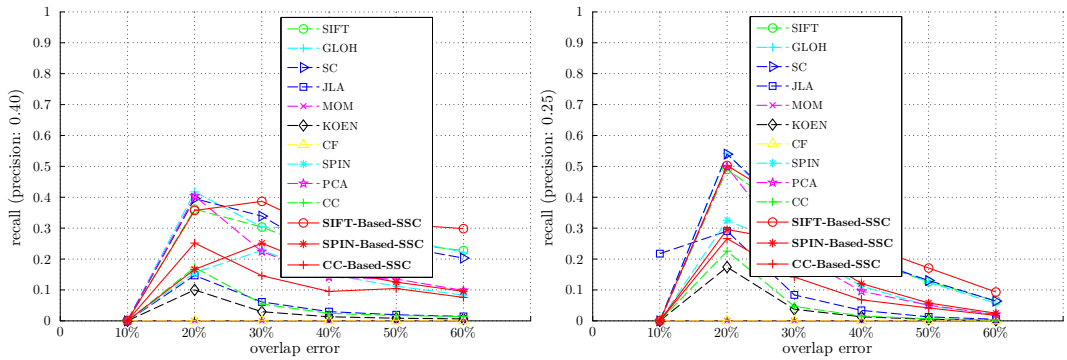
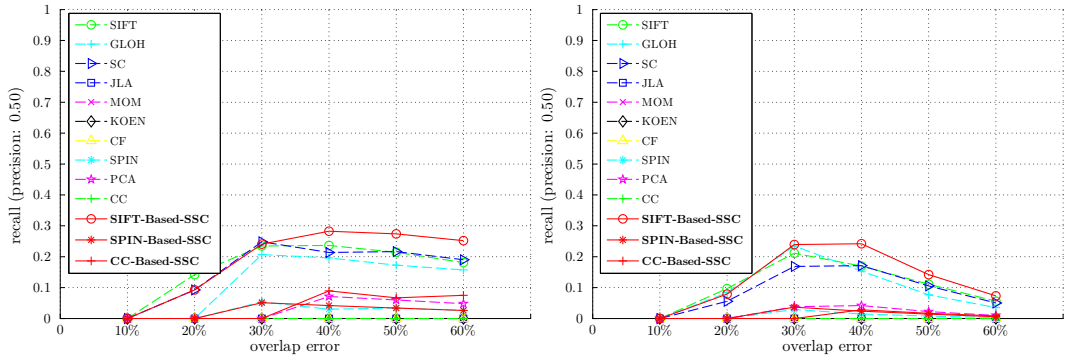


FIG. 5.42: Performance evaluation under viewpoint change for different overlap errors. The results are obtained for the textured scene, *zeriba* of Fig. 5.1k. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed for the precision thresholds of (a) 0.50 and (b) 0.40. The precision scores are computed for the recall thresholds of (c) 0.25 and (d) 0.25.



(a) (b)

FIG. 5.43: Performance evaluation under viewpoint change for different overlap errors. The results are obtained for the structured scene, *graffiti* of Fig. 5.2c. The descriptors are computed for Harris-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed for the precision thresholds of (a) 0.40 and (b) 0.25.



(a) (b)

FIG. 5.44: Performance evaluation under viewpoint change for different overlap errors. The results are obtained for the structured scene, *graffiti*, of Fig. 5.2c. The descriptors are computed for Hessian-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed for the precision thresholds of (a) 0.40 and (b) 0.25.

5.3.6.2 Discriminative power

In these experiments, we assess the discriminative power of the descriptors for viewpoint changes by using three different scenes, `zeriba`, `graffiti`, and `wall`.

For each scene, we use two images to express small and large viewpoint angles. In addition, we consider the affine region detectors of `harris-affine` and `hessian-affine` and matching strategies of `nearest-neighbor` and `threshold-based-matching`. The obtained results are presented in Figs. 5.45-5.50.

In general, we observe from the reported results that the SIFT-Based-SSC still obtains the best discriminative power for both textured and structured scenes. This is specifically for images of large viewpoint angles.

In addition, we notice that the discriminative power of the other SSC-based descriptors are highly increased when including the SSC component. It is also important to consider from Figs. 5.45a, 5.45b, 5.46a, and 5.46b that in the case of the textured scene of `zeriba`, the other approaches (except SSCs) obtain recall scores of ≈ 0 for the *1-precision*-range below ≈ 0.2.

This means that these descriptors are becoming partially unsuitable for large viewpoint changes. However, when they are plugged with the new SSC-based descriptors, they turn into the most suited descriptors while recording high recall scores for high precision range, *i.e.*, *1-precision*-range < 0.2

More than expected, we remark again how well the discriminative power of SSC-based descriptors is enhanced under viewpoint changes even for the standard structured scene of `graffiti` as shown in Figs. 5.47a, 5.47b, 5.48a, and 5.48b.

For further details, we observe from 5.46a and 5.46b that under large viewpoint angles for `zeriba` scene, the effectiveness of SSC approach is vital in increasing the discriminative power of descriptors. Thus, while other descriptors failed by providing low recall scores for high precisions, SIFT-Based-SSC achieves large discriminative power by producing large number of correct matches with high precisions.

Tab. 5.5 below illustrates clearly this case. This table is obtained for a sample of the best descriptors, which are computed for `hessian-affine` regions. The number of correct matches given in this table are computed for the precision scores of 0.1 and 0.4. These are obtained with the `nearest-neighbor` and `threshold-based` matching methods, respectively.

For example, we see in the first row of this table that when integrating the semantic-context information, the numbers of correct matches jump from 0 to 451 for SIFT and attains 193 for CC. This shows how well the SSC influences the performance of

descriptor in presence of large number of similar regions.

TAB. 5.5: The number of correct matches with respect to particular precision thresholds obtained for the scene of *zeriba*. The scores are computed for Hessian-Affine detectors with the precision thresholds of (1st *row*) 0.1 and (2nd *row*) 0.4.

Matching method	SIFT	GLOH	SC	CC	SIFT-Based-SSC	CC-Based-SSC
nearest-neighbor matching	0	32	0	0	**451**	193
Threshold-based matching	0	2759	1379	0	**5173**	2414

More than expected we obtained again SSC-based descriptor performs best in the structured scene of `graffiti`. This is happening particularly with images obtained from large viewpoint angles as shown in Figs. 5.47a, 5.47b, 5.48a and 5.48b.

It seems that for large viewpoint angles, the standard descriptors (like SIFT and GLOH), turn out to be outperformed by the SIFT-Based-SSC. This means that the SSC information not only help to reduce the ambiguity manifested within scenes of similar motifs like textured scenes, but it can also significantly increase the discriminative power in structured scenes.

Therefore, we observe SIFT-Based-SSC continually in the top spot wining largely the remaining descriptors, that is, the gap between SIFT-Based-SSC and the others is significantly large.

Even though it performs less better than for the textured scene of `zeriba`, we notice that SIFT-Based-SSC still ranked first in the textured scene of `wall`, as shown in Figs. 5.49-5.50. More precisely, it improves hugely the discriminative power of SIFT which seems lesser than GLOH and SC ranked 2 and 3 respectively.

The reason that SIFT-Based-SSC acting differently in these scenes although they are both textured, is because of the higher number of similar regions which are spatially close to each other in the scene of `zeriba` than in the `wall`.

The rational here is that scene of `wall` exhibits a larger number of similar motifs but more dispersed in the space which results in degradation of the number of correct matches as we can see by comparing, for example, Figs. 5.45a and 5.49a. Besides, it is also related the local part represented by SIFT which performs lesser in the scene of `wall` than in the `zeriba`.

In conclusion, we realize the vital role of the semantic-context information in enhancing the discriminative power of descriptors computed for images subjected to viewpoint changes. This is observed especially for the textured scene with high probability of region similarity *i.e.*, `zeriba`. The usefulness of SSC approach for structured scenes is illustrated through the scene of `graffiti`, in which the discriminative power is

significantly increased when SIFT-Based-SSC is constructed.

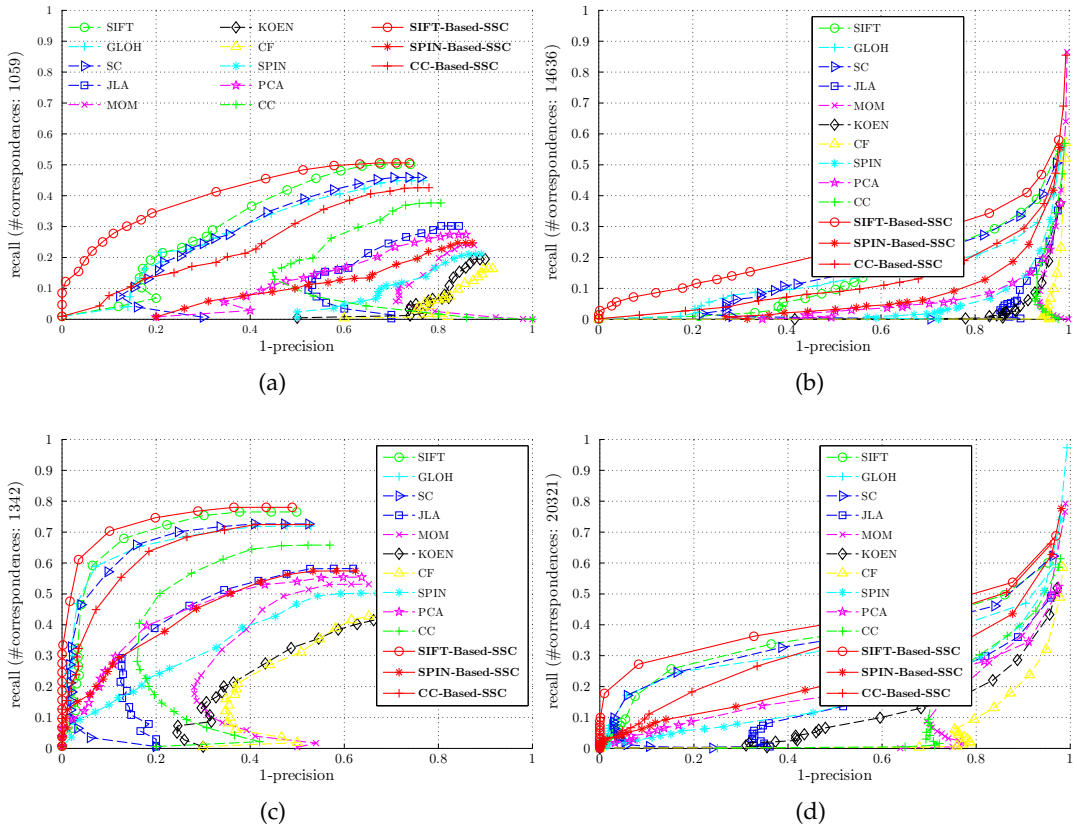


FIG. 5.45: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the textured scene, *zeriba* of Fig. 5.1k. The descriptors are computed for Harris-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 1059, (b) 14636, (c) 1342, and (d) 20321 correspondences.

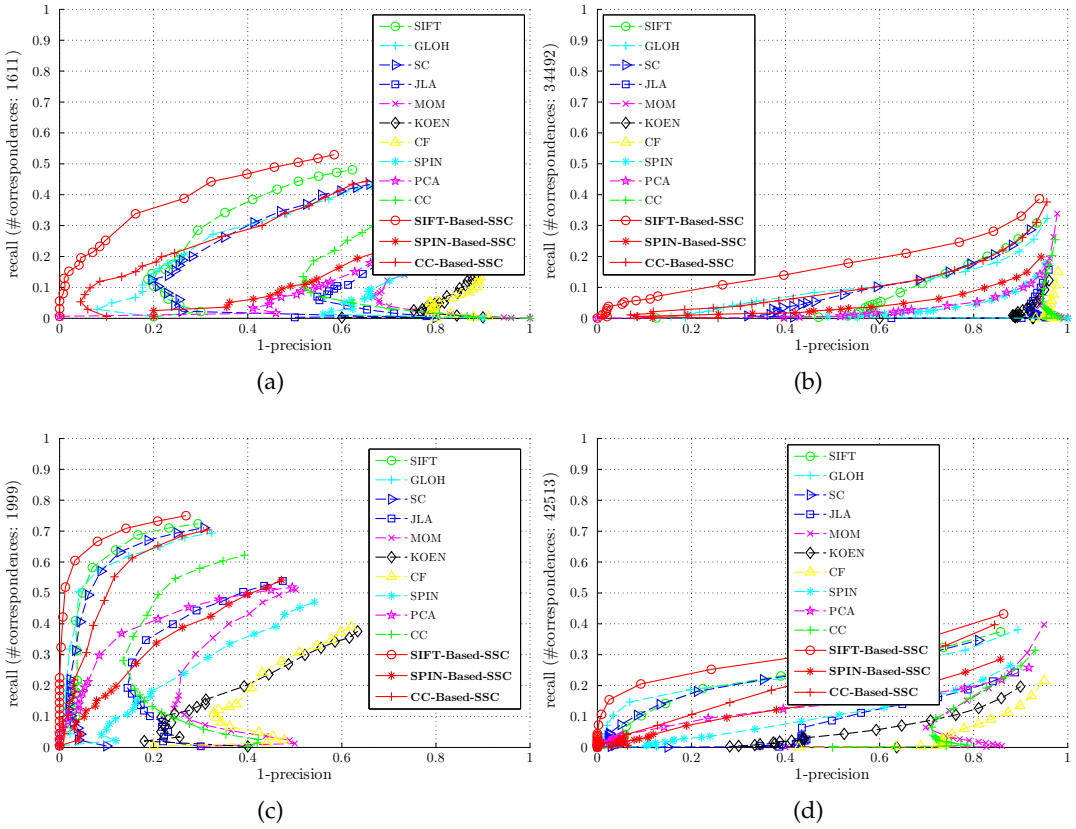


FIG. 5.46: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the textured scene, *zeriba* of Fig. 5.1k. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 1611, (b) 34492, (c) 1999, and (d) 42513 correspondences.

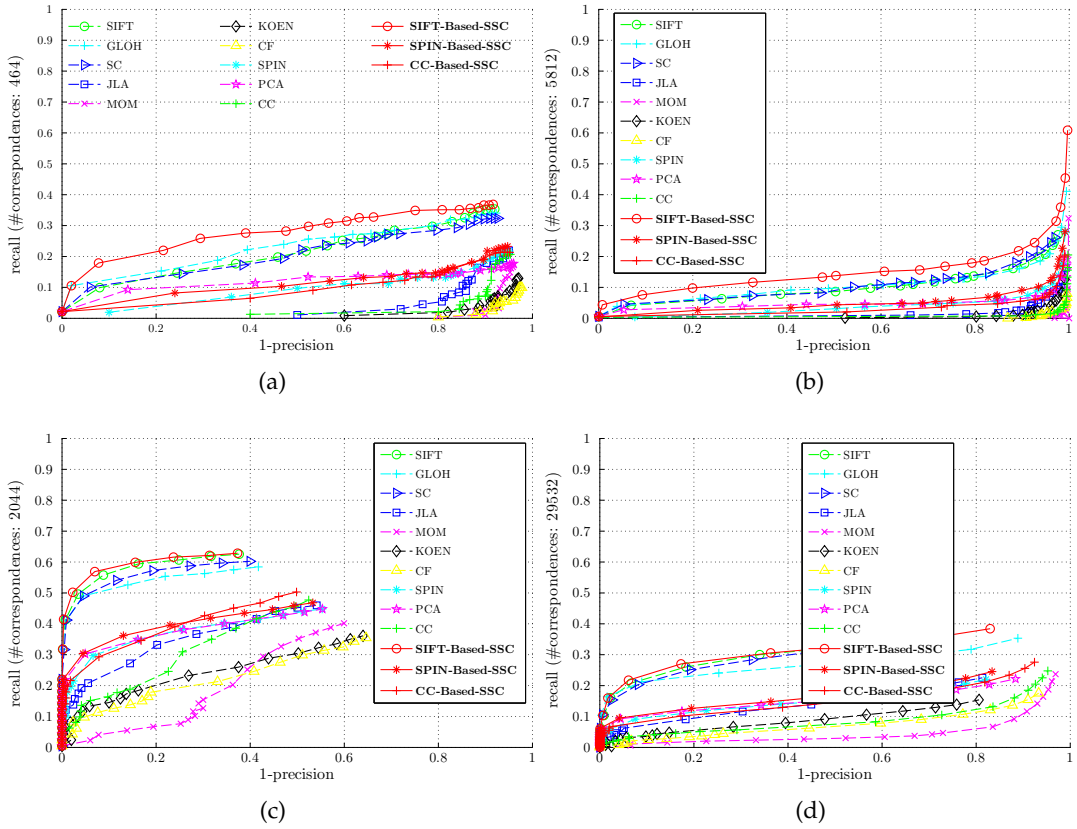


FIG. 5.47: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the structured scene, *graffiti* of Fig. 5.2c. The descriptors are computed for Harris-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 464, (b) 5812, (c) 2044, and (d) 29532 correspondences.

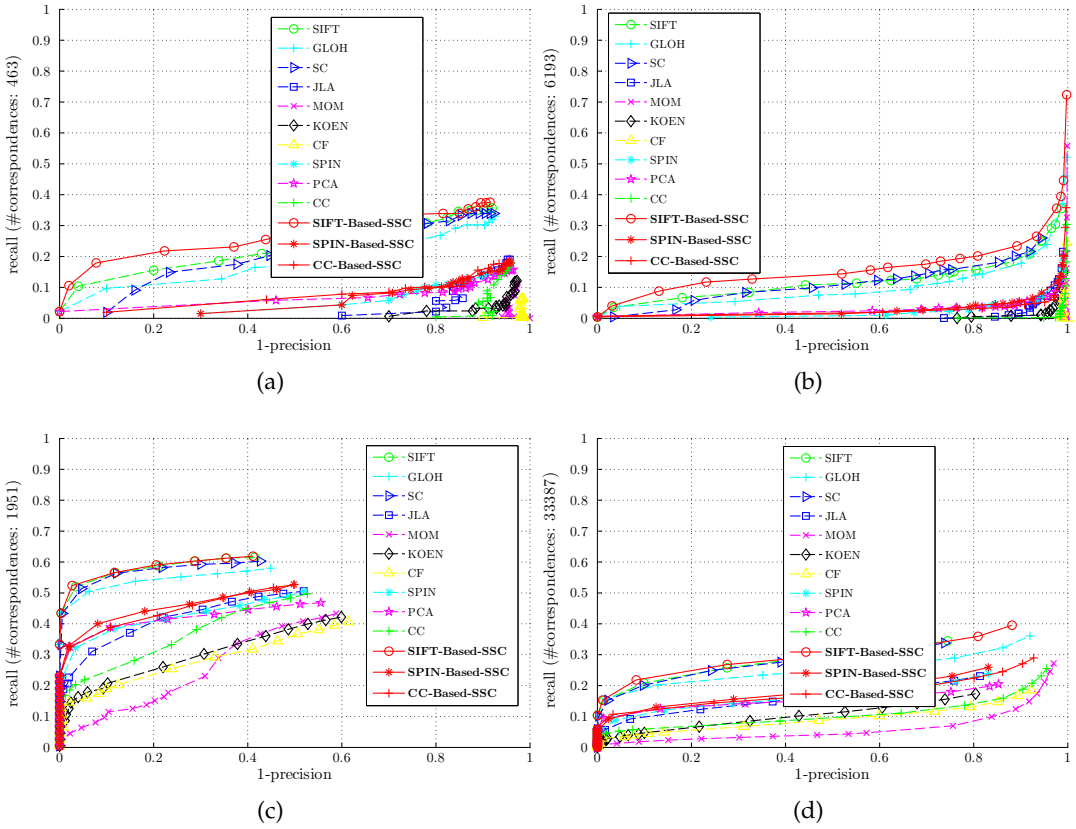


FIG. 5.48: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the structured scene, *graffiti* of Fig. 5.2c. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 463, (b) 6193, (c) 1951, and 33387 (d) correspondences.

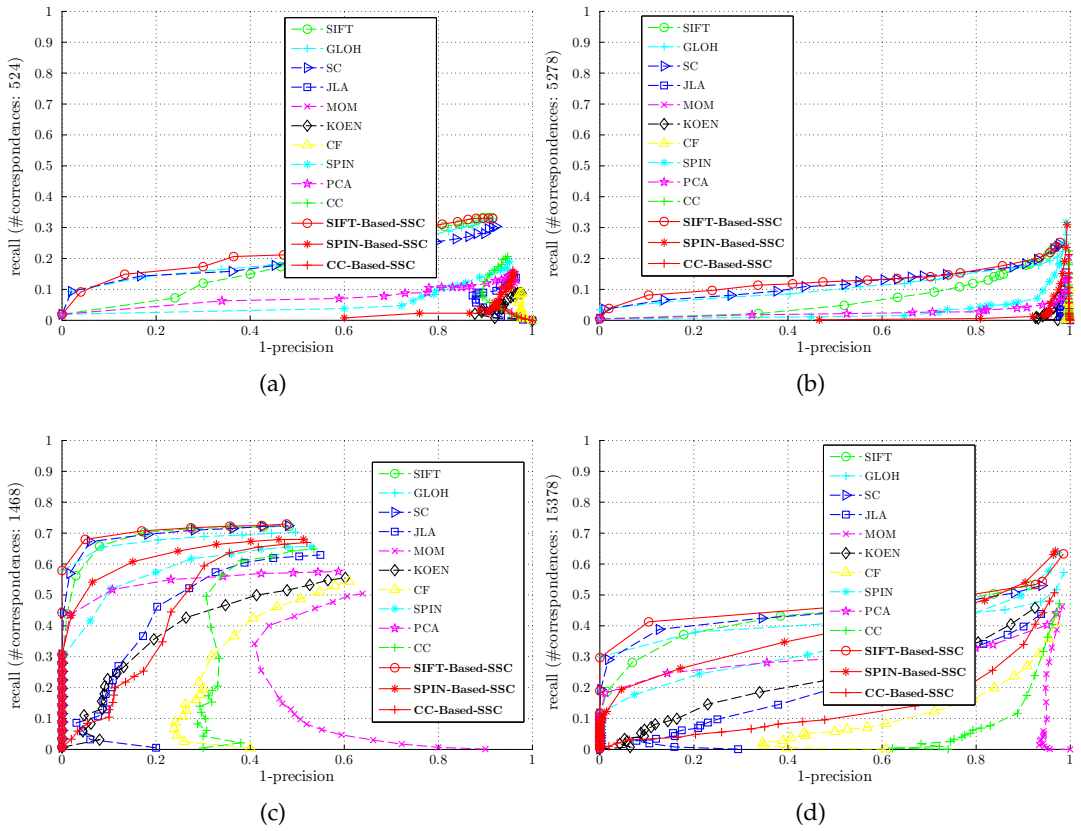


FIG. 5.49: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the scene, *wall* of Fig. 5.2d. The descriptors are computed for Harris-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 524, (b) 5278, (c) 1468, and (d) 15378 correspondences.

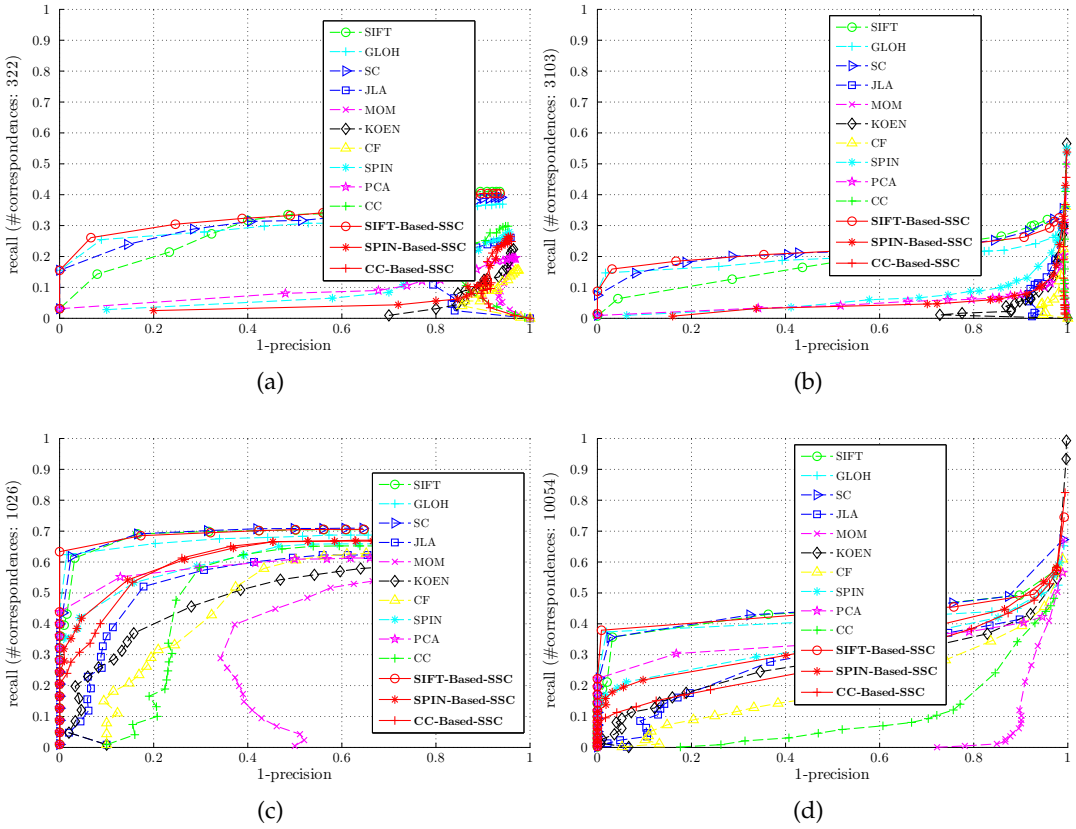


FIG. 5.50: Evaluation results of the discriminative power under viewpoint change. The results are shown for images with (a)(b) large and (c)(d) small viewpoint angles of the textured scene, *wall* of Fig. 5.2d. The descriptors are computed for Hessian-Affine regions and matched using (a)(c) the nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to (a) 322, (b) 3103, (c) 1026, and (d) 10054 correspondences.

5.3.6.3 Invariance

In this evaluation, we investigate the invariance of the SSC-based approach while comparing it to other methods. The investigation is conducted on images subjected to gradual change in viewpoint angles.

These images are those of the textured and structured scenes of `zeriba` and `graffiti`. For each scene, the descriptors are computed for both `harris-affine` and `hessian-affine` region and matched with `nearest-neighbor` and `threshold-based` matching algorithms.

Each descriptor invariance is inspected by measuring the degradations of both recall and precision scores while the viewpoint angle is varied in range of $10° - 70°$ for the scene of `zeriba` and $20° - 60°$ for the scene of `graffiti`.

The recall and precision thresholds are almost higher for the `nearest-neighbor` than for the `threshold-based` matching method. This is because, the ROC curves for the first are mostly close to the left side, whereas they are spread over whole ROC space for the second. The evaluation results are displayed in Figs. 5.51-5.54.

From these results, we discover rapidly the usefulness of adding the SSC component into descriptors like SIFT and CC. It seems the CC invariance is greatly augmented for both recall and precision scores. In this way, SIFT gained the best constancy for both scenes within all evaluations. We also observe that much more gains are obtained, as expected, for the textured scene of `zeriba`.

The results obtained on this latter scene, show SIFT-Based-SSC curves being permanently above those obtained with the other descriptors while coming down more slowly than others. This situation can be observed, for example, in Fig. 5.51a in which while the SIFT-Based-SSC recall falls to 0.3, the others like GLOH, SC, and SIFT drop to 0.15, 0.11, and 0.09, respectively.

Similar observation can be done for the precision curves, as shown, for example, in the Fig. 5.51c. Thus, for the viewpoint angle of 70°, we read a precision of 0.59 for SIFT-Based-SSC and 0.45, 0.27, 0.22 for SIFT, GLOH and SC, respectively.

By comparing the curves of CC obtained in Figs. 5.52c and 5.52d, before and after plugging the SSC part, *i.e.*, CC-Based-SSC, we realize how well the precision of CC is positvely affected when considering SSC information, especially for large viewpoint angles.

For instance, for a viewpoint angle of 50°, as shown in Fig. 5.52c, while the CC gives a precision of zero, the CC-Based-SSC attains ≈ 0.42 and becomes thus competitive to GLOH and SC.

For the structured scene of `graffiti`, the results are slightly different from those obtained in the textured scene of `zeriba`. Though SIFT-Based-SSC being ranked as the best, the discrepancy between it and the others becomes smaller. Tab. 5.6 compares the disparities in terms of the recall and precision scores between the textured and structured scenes.

These are computed for viewpoint angles using SIFT-Based-SSC and the second best, (*i.e.*ranked) descriptor. The recall and precision scores are calculated for the precision and recall thresholds of 0.70 and 0.20, respectively. These scores are obtained with the descriptors computed for `hessian-affine` regions and matched with the `threshold-based` matching algorithm.

TAB. 5.6: Comparison of disparities in the (a) recall and (b) precision scores between SIFT-Based-SSC and the second ranked descriptor. These are obtained for both textured and structured scenes of *zeriba* and *graffiti*, respectively. The disparities are calculated for (a) a precision of 0.70 and (b) a recall of 0.20. Descriptors are computed for Hessian-Affine detector and matched with the threshold-based matching algorithm.

(a) Recall

	20°	30°	40°	50°	60°
Zeriba	**0.02**	**0.03**	**0.05**	**0.07**	**0.08**
Graffiti	0.00	0.01	0.02	0.03	0.04

(b) Precision

	20°	30°	40°	50°	60°
Zeriba	**0.07**	**0.15**	**0.23**	**0.25**	**0.25**
Graffiti	0.00	0.02	0.04	0.20	0.09

This table shows clearly how much the discriminative power, *i.e.*, recall and precision, is increased when SIFT-Based-SSC is adopted, in the textured scene in particular. This appears more straightforward for large viewpoint angles, in which the SIFT-Based-SSC performs much better while the other descriptors are constrained to produce enough recall scores with high precisions.

In conclusion, these experiments show the effectiveness of the SSC component in increasing both discriminative power and invariance of descriptors computed on images subjected to the most complicated non-affine geometric transformation of out-of-plane rotation, or viewpoint change.

In addition, the important performance is obtained by SIFT-Based-SSC for the textured scene under large viewpoint angles.

More than we expected, SIFT-Based-SSC is found also to perform best even for the structured scene with a reduced number of similar motifs, like that of the `graffiti` obtained from the standard data set of *Mikolajczyk*.

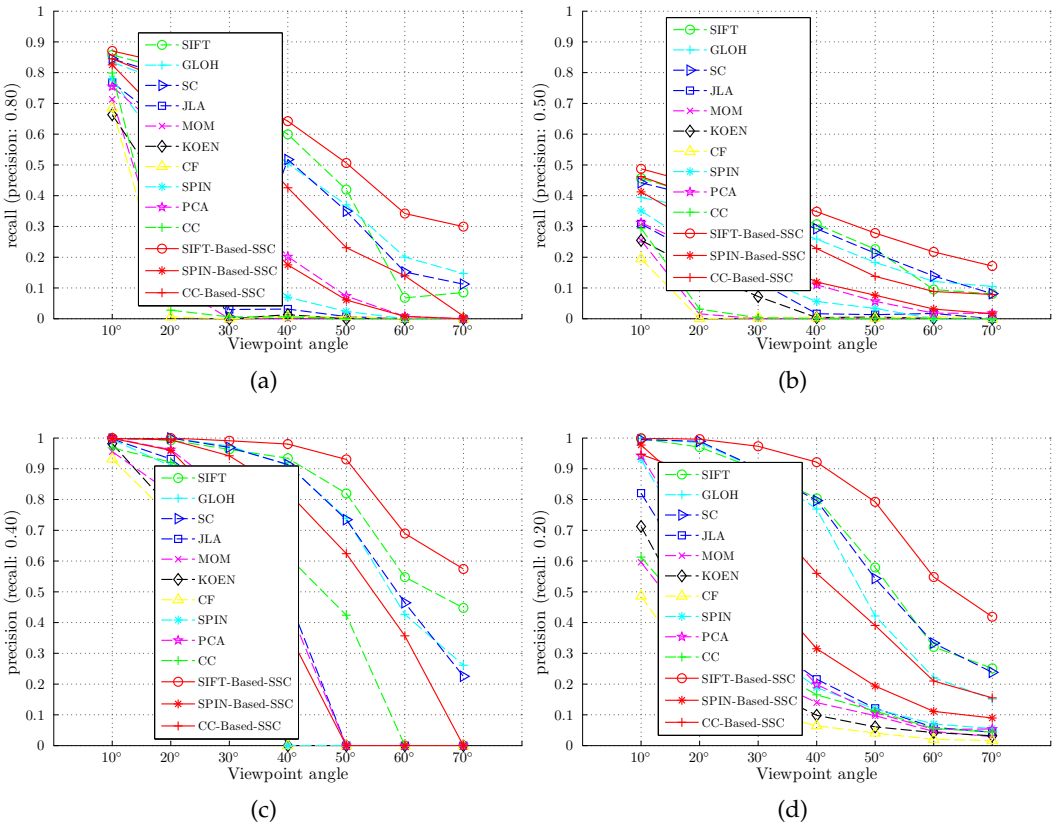


FIG. 5.51: Evaluation results of the invariance under viewpoint changes. The results are obtained for the textured scene, *zeriba* of Fig. 5.1k. The descriptors are computed for Harris-Affine regions and matched using (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.80 and (b) 0.50. The precision scores are computed with respect to the recall thresholds of (c) 0.40 and (d) 0.20.

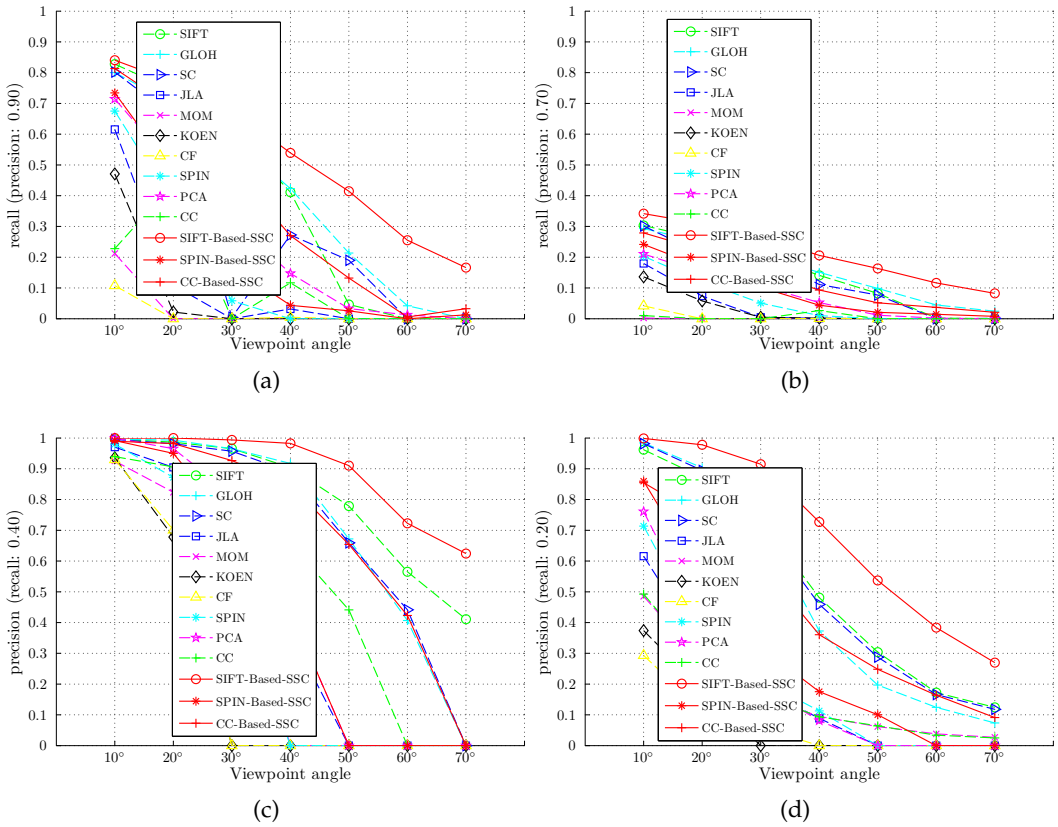


FIG. 5.52: Evaluation results of the invariance under viewpoint changes. The results are obtained for the textured scene, zeriba of Fig. 5.1k. The descriptors are computed for Hessian-Affine regions and matched with (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.90 and (b) 0.70. The precision scores are computed with respect to the recall thresholds of (c) 0.40 and (d) 0.20.

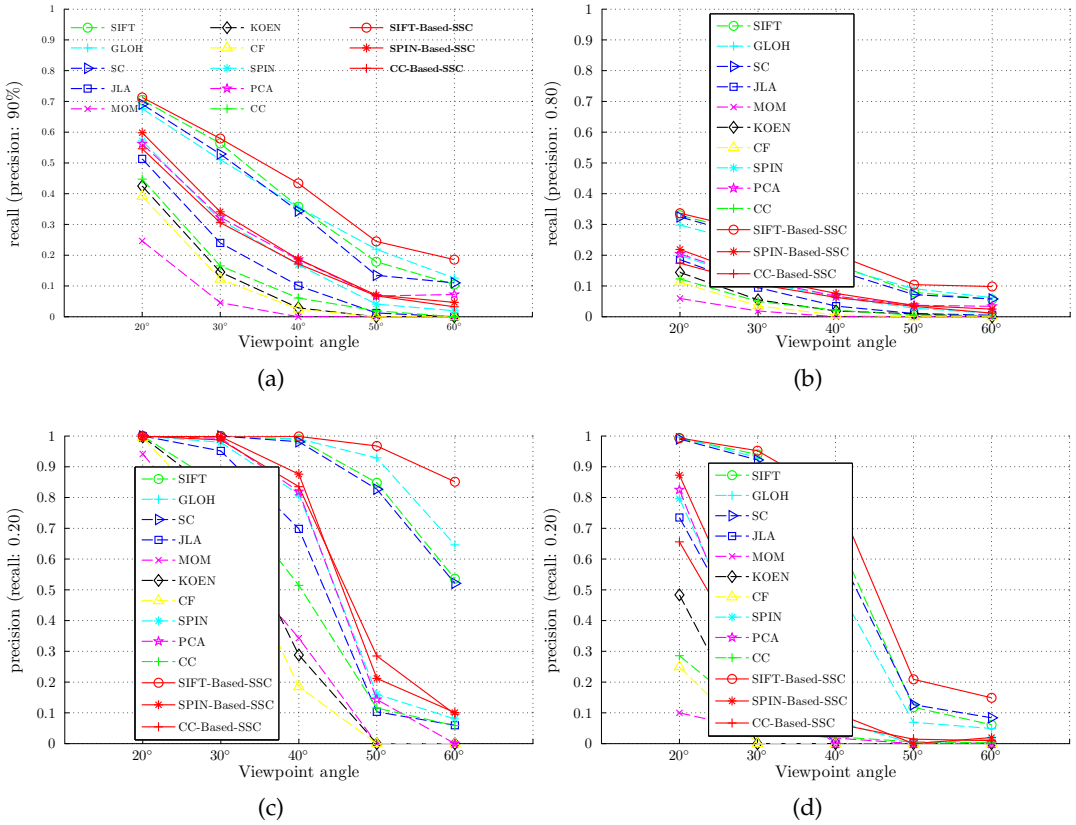


FIG. 5.53: Evaluation results of the invariance under viewpoint changes. The results are shown for the structured scene, *graffiti* of Fig. 5.2c. The descriptors are computed for Harris-Affine regions and matched with (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.90 and (b) 0.80. The precision scores are computed with respect to the recall thresholds of (c) 0.20 and (d) 0.20.

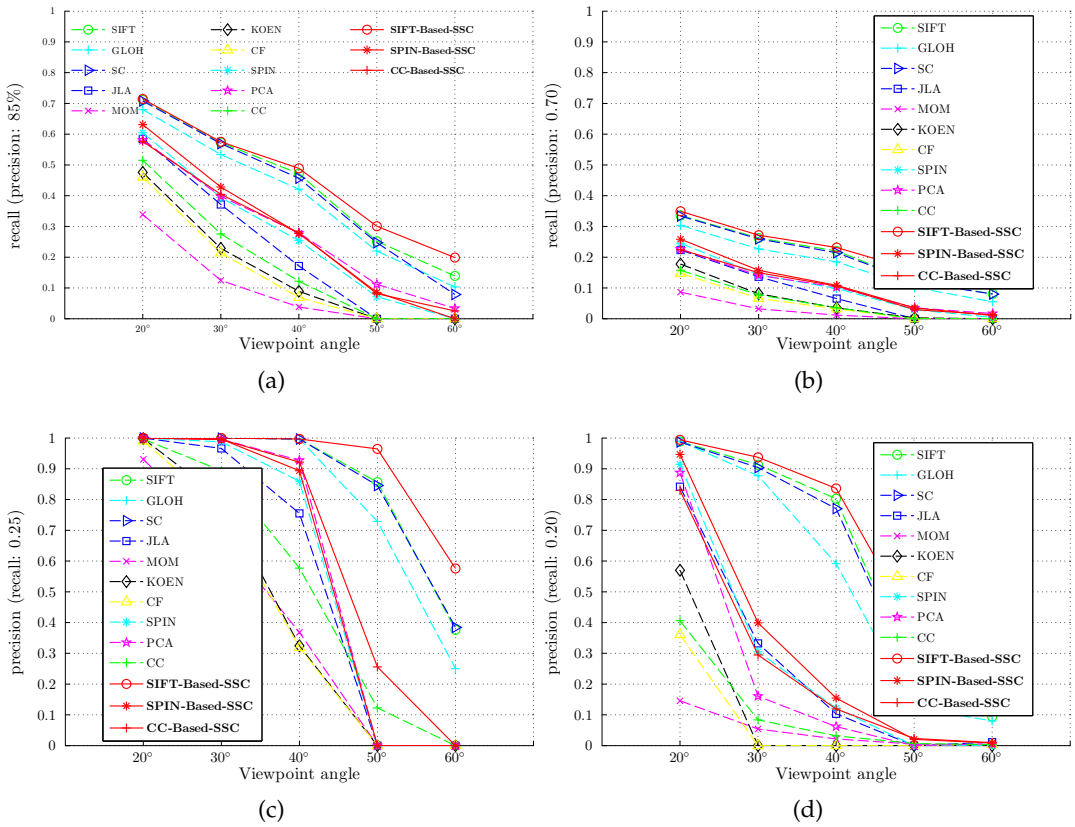


FIG. 5.54: Evaluation results of the invariance under viewpoint changes. The results are shown for the structured scene, *graffiti* of Fig. 5.2c. The descriptors are computed for Hessian-Affine regions and matched with (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision thresholds of (a) 0.85 and (b) 0.70. The precision scores are computed with respect to the recall thresholds of (c) 0.25 and (d) 0.20.

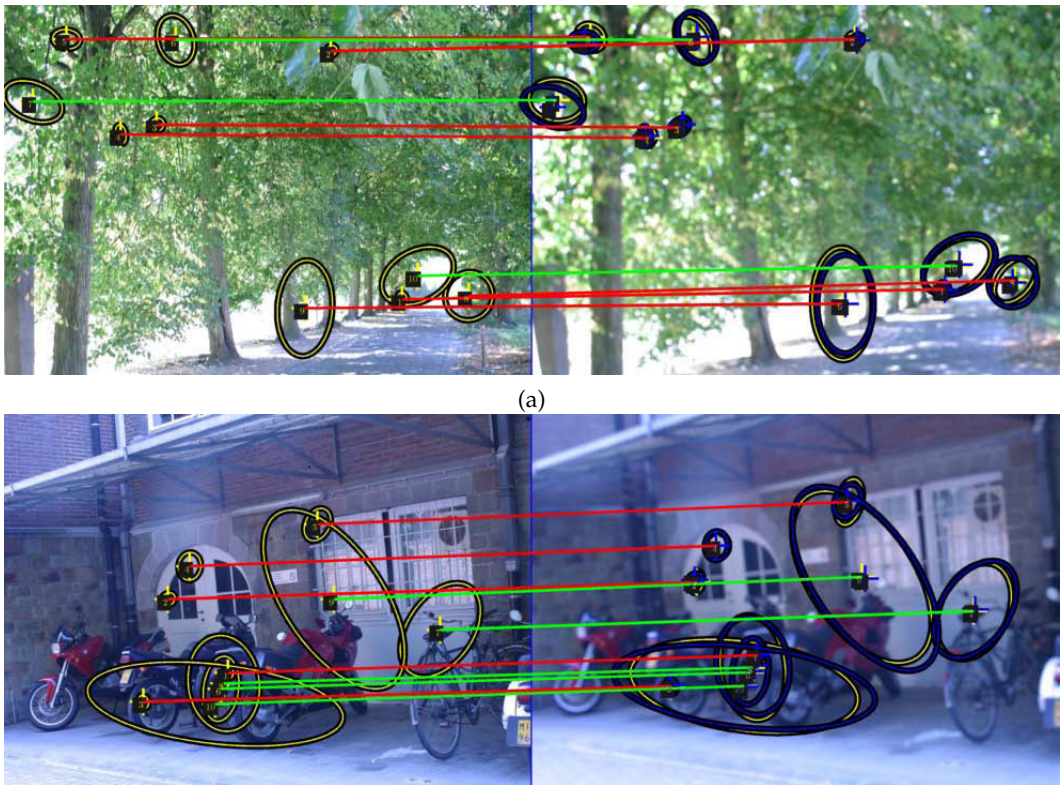
5.3.7 Image Blur

For blur imaging condition, the descriptor performances are evaluated with two different scenes obtained from `Mikolajczyk`'s dataset. The first is the textured scene of `trees` shown in Fig. 5.2f, while the second is the structured of `bikes` shown in Fig. 5.2e.

The goal of this experiment is inspecting the degradation in the discriminative power and invariance when the imaging condition resulting in adding some amount of blur in images, such case for example, variation of the camera focus.

For the discriminative power evaluation, the descriptors are tested using different region detectors and matching strategies. Regarding the invariance evaluation, only `harris-affine` is used with different matching algorithms.

An example of the `nearest-neighbor` matching using SIFT-Based-SSC descriptor computed for `harris-affine` is highlighted in Fig. 5.55.



(a)

(b)

FIG. 5.55: An example of nearest-neighbor matching using SIFT-Based-SSC descriptor computed for Harris-Affine regions. The results are for the (a) textured (*trees*) and (b) structured (*bikes*) scenes. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines.
For the purpose of clarity, only few correspondences are shown.

5.3.7.1 Discriminative power

Here, the discriminative power of the descriptors are chekced using both scenes of `trees` and `bikes`. For each scene, the descriptors are computed on regions extracted by `harris-laplace`, `hessian-laplace`, `harris-affine` and `hessian-affine` detectors. These are then matched with `nearest-neighbor` and `threshold-based` matching algorithms. The obtained results are highlighted in Figs. 5.56-5.63.

Overall, we observe that the discriminative powers of SIFT, CC and SPIN descriptors are significantly improved after adding semantic-context informations resulting in bringing, as example, SIFT-Based-SSC to the top rank in all evaluations.

Yet more in the structured scene, for which in some cases SIFT is outperformed by other descriptors like GLOH, PCA-SIFT, etc. We also notice the large impact of the SSC information in increasing the discriminative power of the CC descriptor as we can see, as example, in Fig. 5.63.

More precisely, we remark that for the textured scene, `trees`, the ranking of descriptors is almost the same within all evaluations. That is, for different region detectors and matching strategies. This places SIFT-Based-SSC first followed by SIFT, SC, GLOH, PCA-SIFT in the 2nd, 3rd, 4th, and 5th ranks, respectively.

Furthermore, we observe that the gaps between SIFT-Based-SSC and other descriptors become more important inside regions of highest precisions, *i.e.*, ranges of 1-precision below than ≈ 0.1 as noticed in Figs. 5.56-5.59.

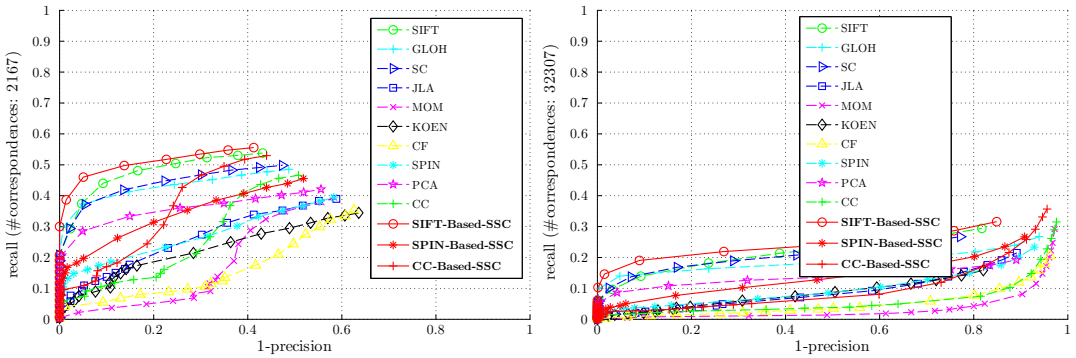
These regions correspond to the left-hand sides of the ROC graphs, which are more revealing of the descriptor discriminative power. In addition, we notice that the discriminative power of CC is also well improved with CC-Based-SSC.

The curves obtained in Figs. 5.60-5.63, which are related to the structured scene of `bikes`, show SIFT-Based-SSC outperforming other descriptors for all region detectors and matching algorithms.

Moreover, we observe that the discriminative power of both CC and SPIN are clearly reinforced when they are substituted by CC-Based-SSC and SPIN-Based-SSC, respectively. For instance, the left-hand side (more interesting region in ROC space) of Fig. 5.63a shows PCA-SIFT and JLA becoming outperformed by CC when adding SSC component. That is to say, the CC curve is below those of PCA-SIFT and JLA which in turn are farther down the curve of CC-Based-SSC.

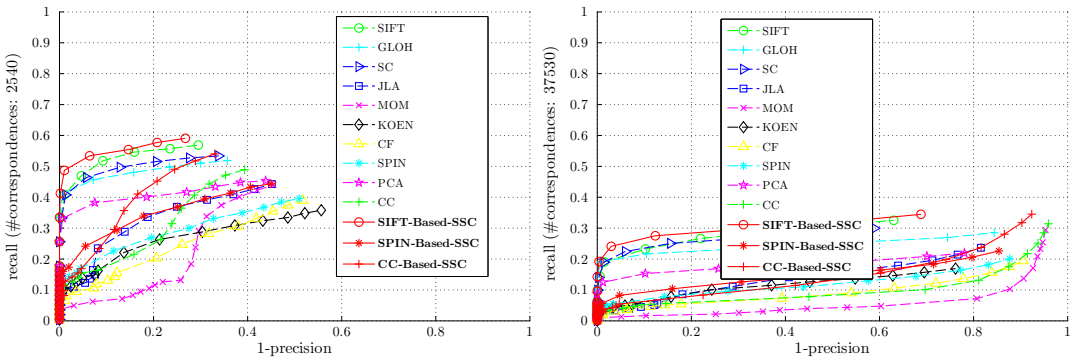
In conclusion, we obtained across these evaluations, the SIFT-Based-SSC performs the best in the discriminative power through all experiment and for the textured scene in particular. We have also seen the discriminative power of CC and SPIN significantly

increased when substituting CC-Based-SSC and SPIN-Based-SSC descriptors.



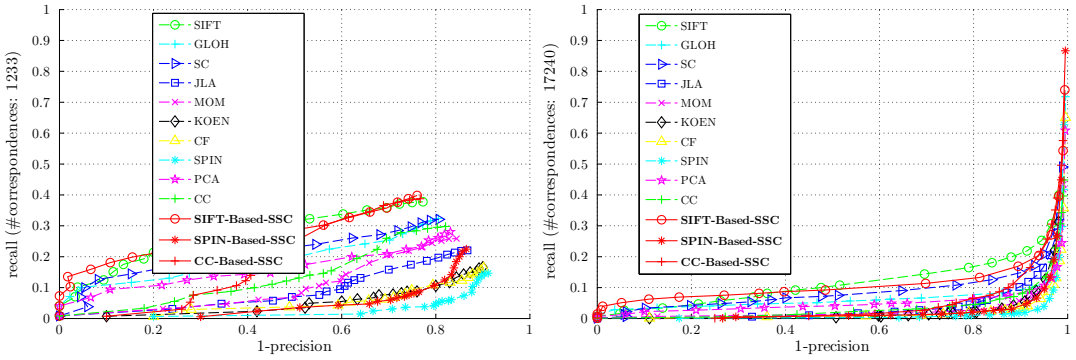
(a) (b)

FIG. 5.56: Evaluation results of the discriminative power under image blur. The results are obtained for the textured scene, *trees* of Fig. 5.2f. The descriptors are computed for Harris-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2167 and (b) 32307 correspondences.



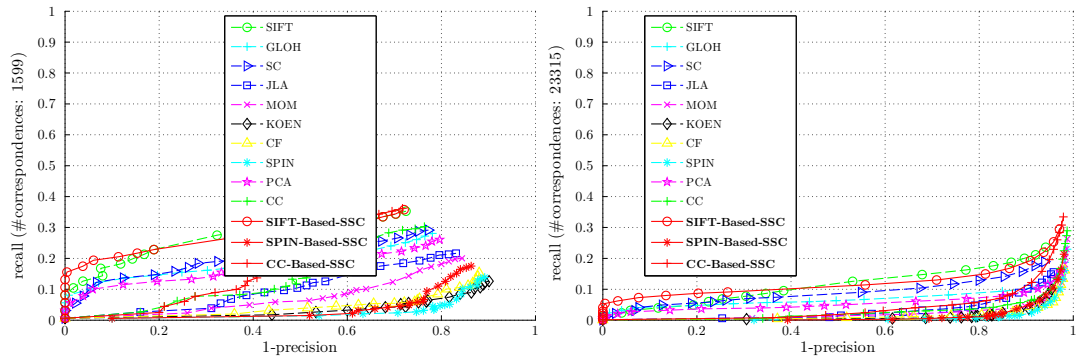
(a) (b)

FIG. 5.57: Evaluation results of the discriminative power under image blur. The results are obtained for the textured scene, trees of Fig. 5.2f. The descriptors are computed for Hessian-Laplace regions and matched using the (a) nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are with respect to (a) 2540 and (b) 37530 correspondences.



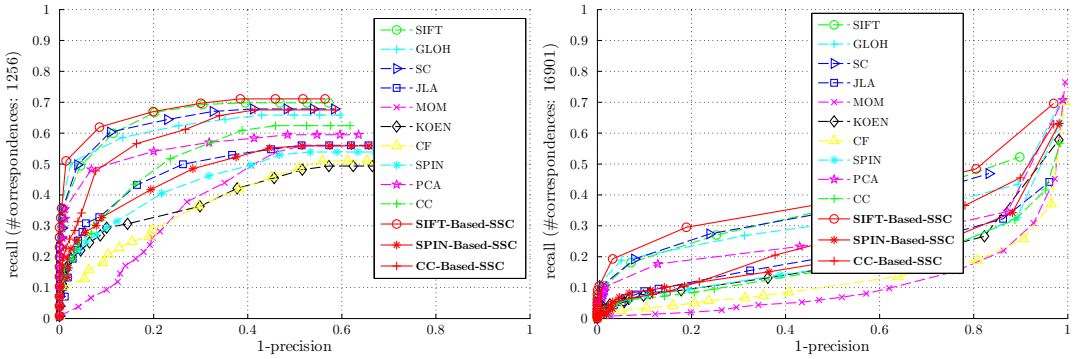
(a) (b)

FIG. 5.58: Evaluation results of the discriminative power under image blur. The results are obtained for the textured scene of *trees* of Fig. 5.2f. The descriptors are computed for Harris-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1233 and (b) 17240 correspondences.



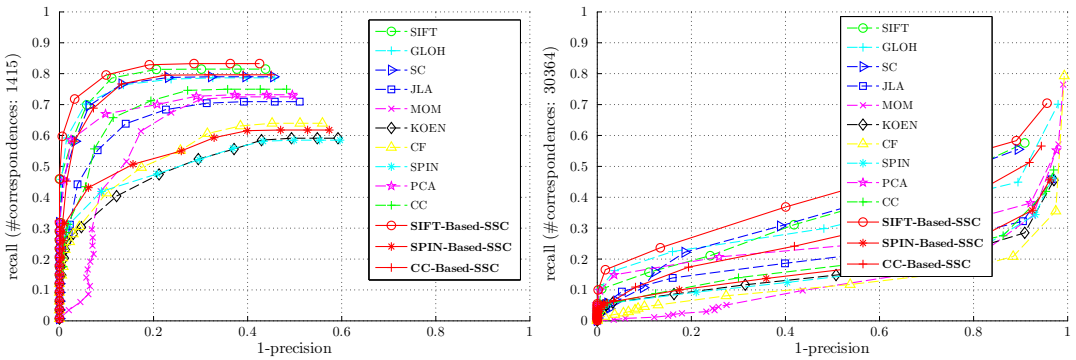
(a) (b)

FIG. 5.59: Evaluation results of the discriminative power under image blur. The results are obtained for the textured scene of *trees* of Fig. 5.2f. The descriptors are computed for Hessian-Affine regions and matched using (a) the nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are computed with respect to (a) 1599 and (b) 23315 correspondences.



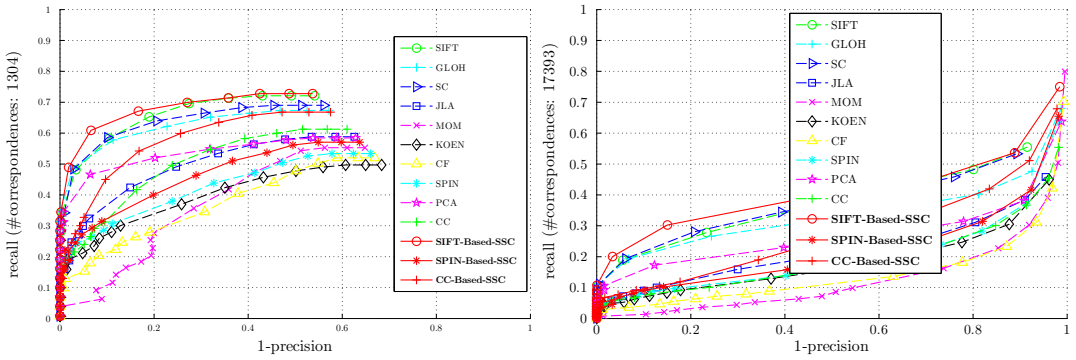
(a) (b)

FIG. 5.60: Evaluation results of the discriminative power under image blur. The results are obtained for the structured scene of *bikes* of Fig. 5.2e. The descriptors are computed for Harris-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1256 and (b) 16901 correspondences.



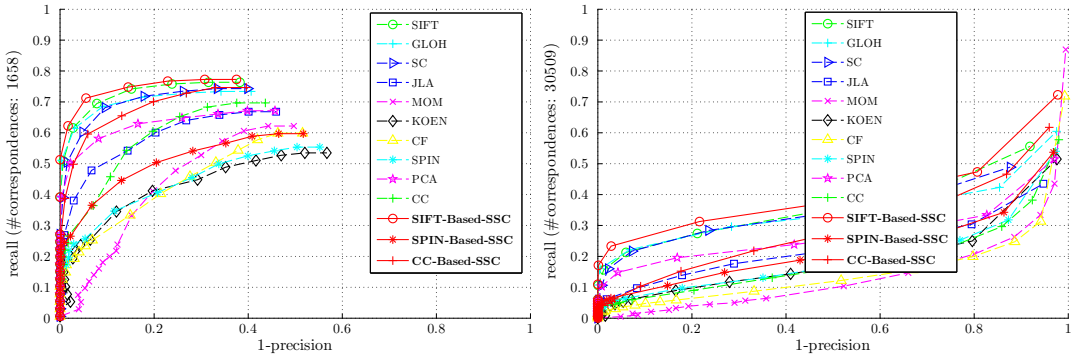
(a) (b)

FIG. 5.61: Evaluation results of the discriminative power under image blur. The results are obtained for the structured scene of *bikes* of Fig. 5.2e. The descriptors are computed for Hessian-Laplace regions and matched using the (a) nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are with respect to (a) 1415 and (b) 30364 correspondences.



(a) (b)

FIG. 5.62: Evaluation results of the discriminative power under image blur. The results are obtained for the structured scene of *bikes* of Fig. 5.2e. The descriptors are computed for Harris-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1304 and (b) 17393 correspondences.



(a) (b)

FIG. 5.63: Evaluation results of the discriminative power under image blur. The results are obtained for the structured scene of *bikes* of Fig. 5.2e. The descriptors are computed for Hessian-Affine regions and matched using (a) the nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are computed with respect to (a) 1658 and (b) 30509 correspondences.

5.3.7.2 Invariance

In this experiment, the invariance of descriptors are evaluated against image blur using textured and structured scenes of `trees` and `bikes`, respectively. For each scene, the invariance is measured by inspecting the degradation in both recall and precision scores while the amount of blur is progressively increasing. This is performed by using range of 5 images in each scene, depicting a gradual augmentation in blur.

The descriptors for each image are computed on support regions obtained by `harris-affine` detector and matched with both `nearest-neighbor` and `threshold-based` matching methods. The obtained results are depicted in Figs. 5.64 and 5.65.

These figures demonstrate, the recall and precision scores obtained with SIFT-Based-SSC degrade more slowly than with the others. Also we observe that with respect to the recall, the invariance of SIFT-Based-SSC is better in the textured scene, as shown in Figs. 5.64a and 5.64b, than in the structured scene as shown in Figs. 5.65a and 5.65b.

However in terms of precision, SIFT-Based-SSC performs much better in the structured scene than in the textured scene. This can be noticed after comparing curves of Figs. 5.64c and 5.64d to those of Figs. 5.65c and 5.65d. This also shows the large gap in precisions between SIFT-Based-SSC and other descriptors for higher amount of blur.

Tab. 5.7 compares the precision scores for the five top best descriptors displayed in Fig. 5.65d. These are computed for the structured scene of `bikes`. The scores are obtained with images of lowest and highest amounts of blur.

TAB. 5.7: Degradations in precision scores under image blur computed for images of the lowest and highest amounts of blur, and with respect to recall of 0.30. The scores are obtained for descriptors computed for Harris-Affine regions on the scene of *bikes* and matched with the threshold-based matching technique.

	SIFT	GLOH	SC	SIFT-Based-SSC	CC-Based-SSC
Lowest blur	0.97	**0.98**	0.97	**0.98**	0.78
Highest blur	0.70	0.60	0.74	**0.86**	0.40

This table shows that for low amounts of blur, most of descriptors except CC-Based-SSC, appear to have approximately similar precisions, whereas when the amount of image blur is highly increased, the precisions of the other descriptors are rapidly dropping down while SIFT-Based-SSC keeps significantly a higher precision.

In conclusion, from above evaluations we have seen that the invariance of descriptors computed for blurred images, can be significantly enhanced when adding SSC information. We also found that with respect to the recall, the impact of the SSC information is better in the textured than in the structured scenes, whereas for the precision, it is the

contrary, better in the structured than in the textured.

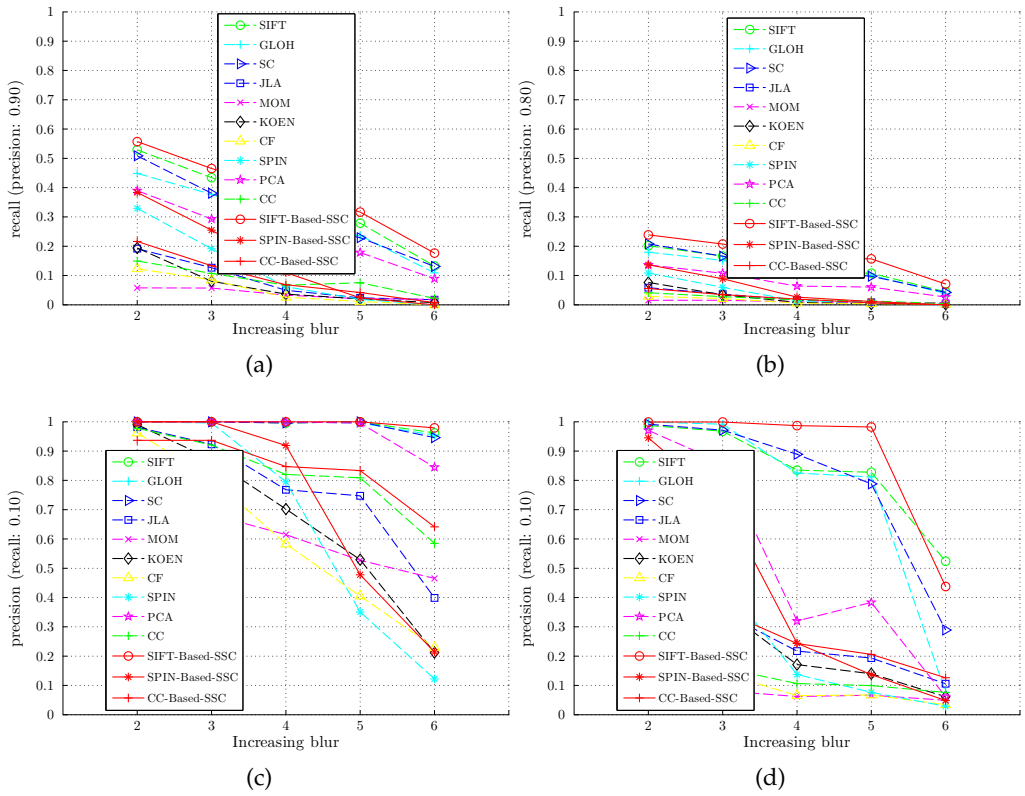


FIG. 5.64: Evaluation results of the invariance under image blur. The results are obtained for the textured scene of *trees* of Fig. 5.2f. The descriptors are computed for Harris-Affine regions and matched using the (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to precision thresholds of (a) 0.90 and (b) 0.80. The precision scores are computed with respect to the recall threshold of 0.10 .

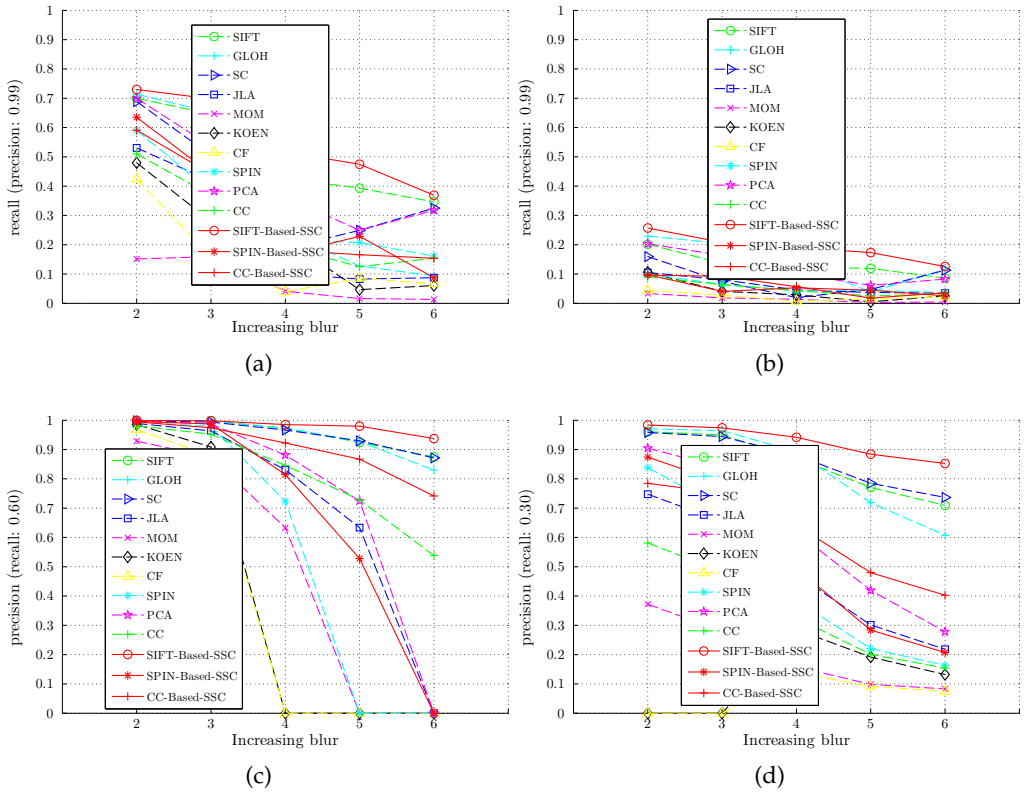


FIG. 5.65: Evaluation results of the invariance under image blur. The results are obtained for the structured scene of *bikes* of Fig. 5.2e. The descriptors are computed for Harris-Affine regions and matched using the (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to the precision threshold of 0.99. The precision scores are computed with respect to recall thresholds of (c) 0.60 and (d) 0.30.

5.3.8 Illumination Changes

The following experiment will report the evaluation of the descriptor performances under illumination changes. The performance is investigated based on the discriminative power and invariance criteria, using the structured scene of `cars` shown in Fig. 5.2g.

This is composed of 6 images reflecting gradual decreases in image lighting. With respect to the discriminative power evaluation, the descriptors are computed on different support regions and then matched with different matching algorithms as well.

For the invariance evaluation, the descriptors are computed for `harris-affine` regions and tested with different matching techniques.

As example, Fig. 5.66 displays the result of `nearest-neighbor` matching obtained for SIFT-Based-SSC descriptor computed on `harris-affine` regions.

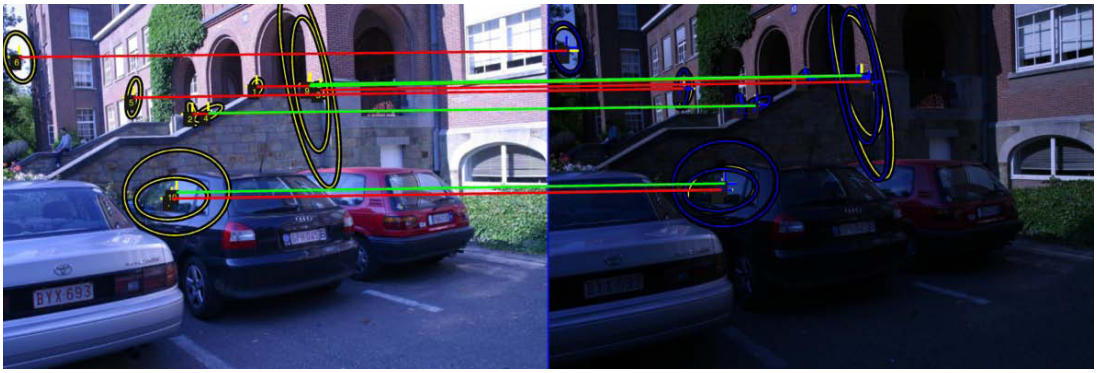


FIG. 5.66: An example of nearest-neighbor matching using SIFT-Based-SSC descriptor computed for Harris-Affine regions. The results are for the structured scene, cars of Fig. 5.2g. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines. For the purpose of clarity, only few correspondences are shown.

5.3.8.1 Discriminative power

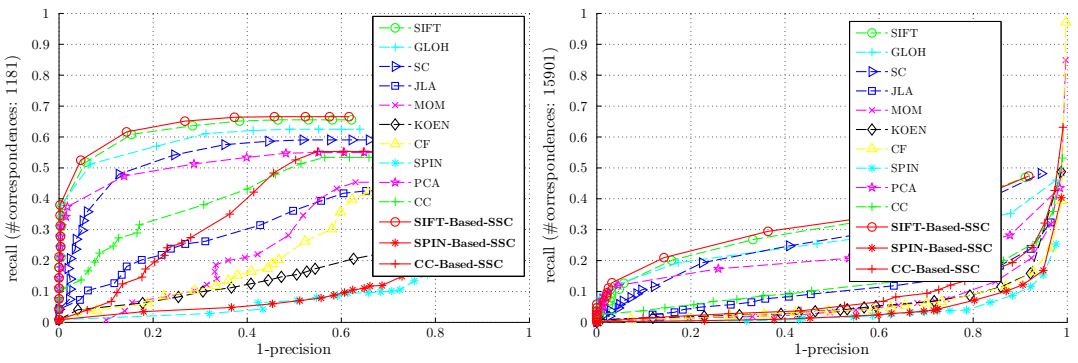
The discriminative power of the descriptors under illumination change are evaluated for different support regions and matching strategies. The results of these evaluations are given in Figs. 5.67-5.70.

These results demonstrate the minor impact of the SSC information in improving the descriptor discriminative power. It is clearly less important than for previous evaluations, *e.g.*, viewpoint change, image blur, etc.

This can be observed for the SIFT-Based-SSC curves across all reported figures. Unfortunately, these figures show also the bad impact of the SSC component when it is added into SPIN and CC descriptors.

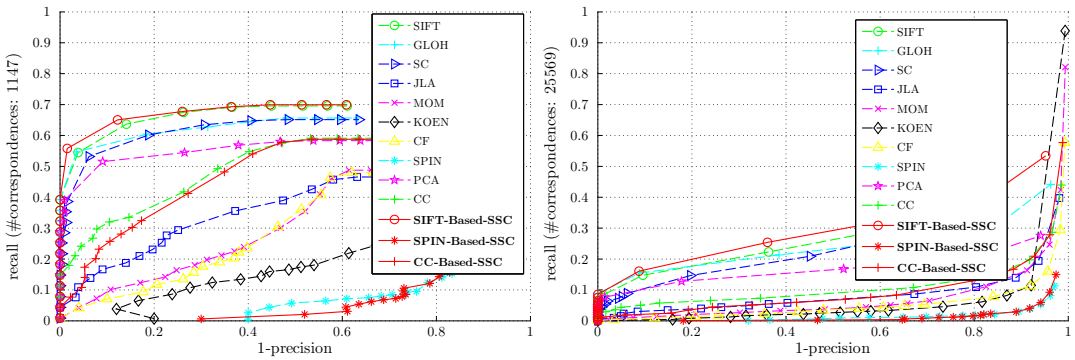
The resulting SPIN-Based-SSC and CC-Based-SSC obtained the discriminative power dramatically decreased, for example, as shown in Fig. 5.69 and 5.70.

Although it performs less better than for other scenes, we observe SIFT-Based-SSC still outperforms the rest of descriptors and ranked permanently in the top spot in all evaluations. Moreover, its highest recall scores are obtained within most interesting regions in ROC space *e.g.*, Fig. 5.68a.



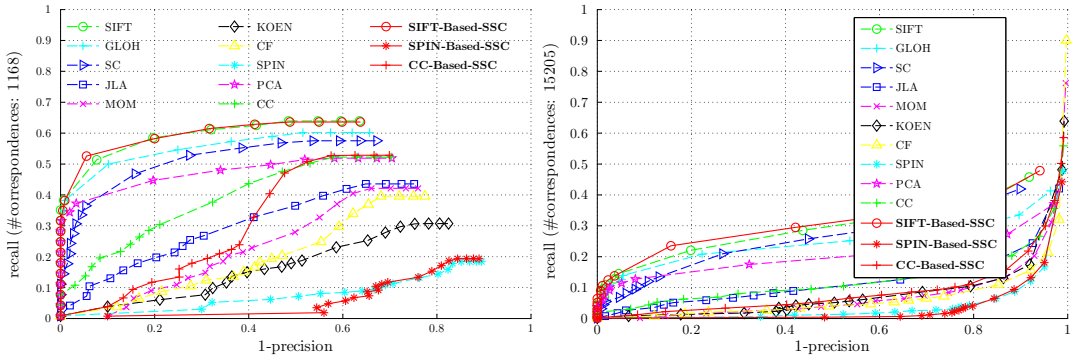
(a) (b)

FIG. 5.67: Evaluation results of the discriminative power under illumination change. The results are obtained for the structured scene of *cars* of Fig. 5.2g. The descriptors are computed for Harris-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1881 and (b) 15901 correspondences.



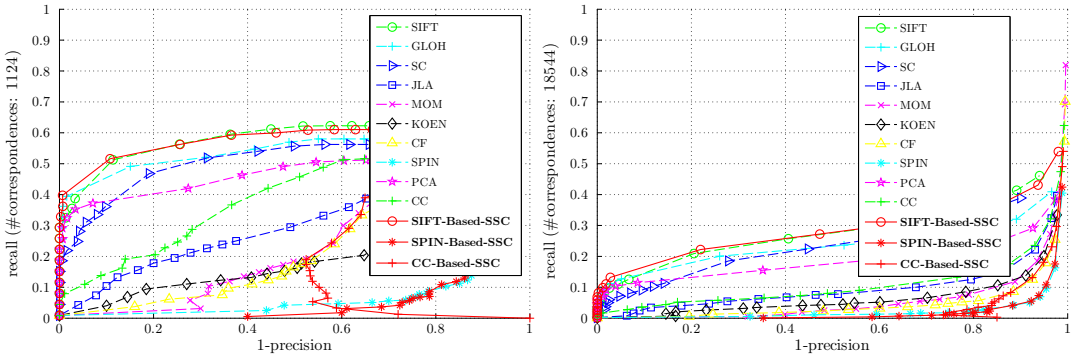
(a) (b)

FIG. 5.68: Evaluation results of the discriminative power under illumination change. The results are obtained for the structured scene of *cars* of Fig. 5.2g. The descriptors are computed for Hessian-Laplace regions and matched using the (a) nearest-neighbor and (b) distance-ratio-based nearest-neighbor matching techniques. The recall scores are with respect to (a) 1147 and (b) 25569 correspondences.



(a) (b)

FIG. 5.69: Evaluation results of the discriminative power under illumination change. The results are obtained for the structured scene of *cars* of Fig. 5.2g. The descriptors are computed for Harris-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1168 and (b) 15205 correspondences.



(a) (b)

FIG. 5.70: Evaluation results of the discriminative power under illumination change. The results are obtained for the structured scene of *cars* of Fig. 5.2g. The descriptors are computed for Hessian-Affine regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 1124 and (b) 18544 correspondences.

5.3.8.2 Invariance

Regarding to the invariance evaluation, we consider descriptors computed `harris-affine` regions and then matched with `nearest-neighbor` and `threshold-based` matching algorithms.

The invariance is measured by inspecting the degradations in recall and precision scores while the scene is subjected to progressive decreases in illumination. This is given through 5 images constituted by the scene of `cars` viewed under different light intensities. The evaluation results are displayed in Fig. 5.71.

Even though it is not pertinent as in the case of viewpoint change or image blur, we observe the invariance of SIFT is improved for both recall and precision cores, and the SIFT-Based-SSC still ranked first in all evaluations. This is more clear for less illuminated image, as shown in 5.71d.

Based on this figure, Tab. 5.8 is established to compare degradations in precision scores for the the top 5 best descriptors. The degradations are computed as the difference between precision score obtained for 5th and 6th images.

In addition, we report along the first and third rows of this table the precision score obtained for the 5th image and the percentages of the degradation amounts, respectively.

TAB. 5.8: Degradations in precision scores under illumination change for the scene of *cars*. The score are calculated with respect to recall of 0.25. The descriptors are computed for Harris-Affine regions and matched with the threshold-based matching technique. The first row shows the precision scores of each descriptor obtained in the 5th image while the second row displays differences between precision scores obtained in the 5th and 6th images, whereas the degradation percentages are recorded along the third row.

	SIFT	SC	GLOH	MOM	SIFT-Based-SSC
Precision	0.88	0.80	0.80	0.56	**0.92**
Degradation ↓	0.18	0.20	0.32	0.36	**0.14**
	−20%	−25%	−40%	−64%	−**15**%

This table illustrates that the SIFT-Based-SSC registers the best scores in both precision and constancy performance in the sense that it obtains the higher precision, *i.e.*, 0.92, and lower slope, *i.e.*, 0.14.

In contrast with SIFT, we see the invariance of SPIN and CC are deteriorated when integrating the SSC information. This is observed particularly for highly blurred images (*e.g.*, images 5 and 6 in Fig. 5.71a) even though significant improvements are noticed for less blurred images as obtained with the recall curves of SPIN-Based-SSC and CC-Based-SSC of figures 5.71a and 5.71b.

To summarize, during these evaluations we figured out that adding the SSC information to SIFT descriptor has a slight effect to improve the descriptor invariance *i.e.*, less than in the case of other deformations (viewpoint change, image blur, etc). However, the performance of SPIN and CC almost decreases when substituting them by SPIN-Based-SSC and CC-Based-SSC, respectively.

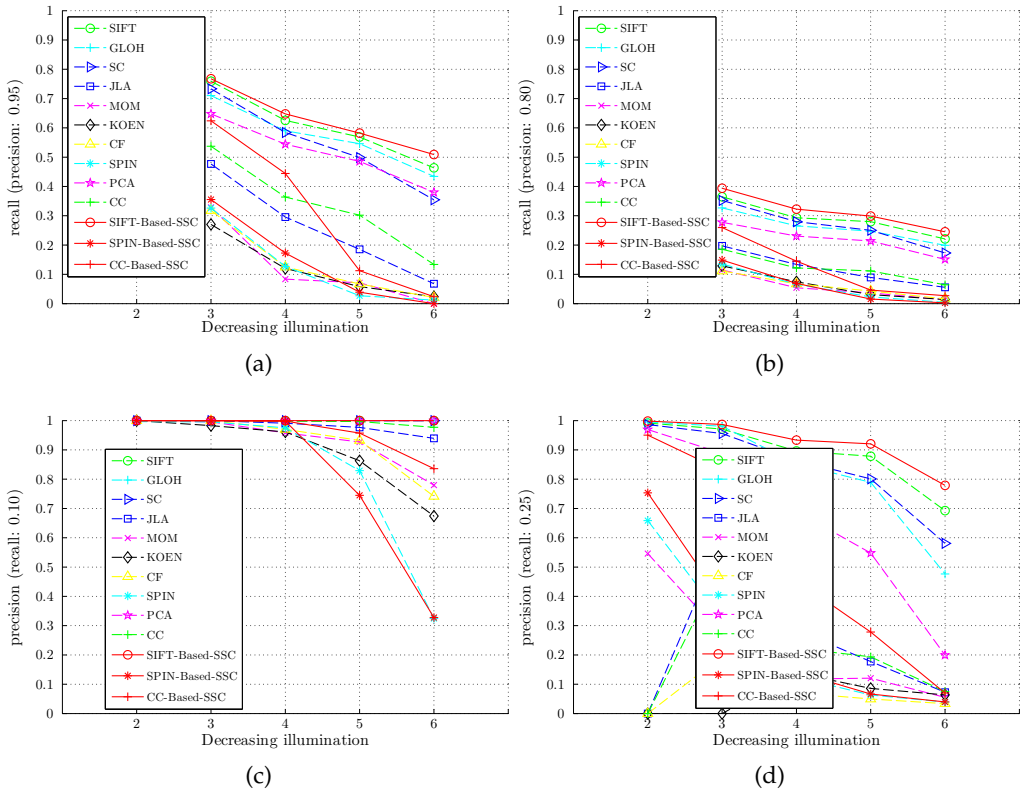


FIG. 5.71: Evaluation results of the invariance under illumination change. The results are obtained for the structured scene of *cars* of Fig. 5.2g. The descriptors are computed for Harris-Affine regions and matched using the (a)(c) nearest-neighbor and (b)(d) threshold-based matching techniques. The recall scores are computed with respect to precision thresholds of (a) 0.95 and (b) 0.80. The precision scores are computed with respect to recall thresholds of (c) 0.10 and (d) 0.25.

5.3.9 JPEG Compression

To evaluate the performance of our approach against loss in image quality, a set of images of different degrees of JPEG-compression are used. These are those of `ubc`'s scene shown in Fig. 5.2h.

The performance are evaluated with respect to the discriminative power of descriptors computed on `harris-laplace` `hessian-laplace`, `harris-affine`, and `hessian-affine` support regions. We adopt both `nearest-neighbor` and `threshold-based` matching approaches.

For instance, Fig. 5.72 shows a sample of support regions (yellow ellipses) of `harris-affine` described using SIFT-Based-SSC descriptor and then matched with `nearest-neighbor` matching technique (green lines).

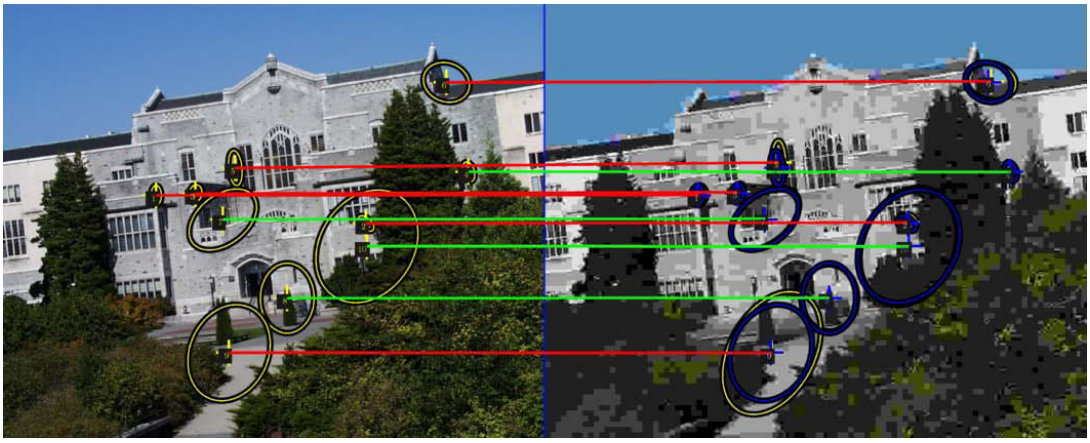


FIG. 5.72: An example of nearest-neighbor matching using SIFT-Based-SSC descriptor computed for Harris-Affine regions. The results are obtained for the structured scene, ubc of Fig. 5.2h. The detected regions are in yellow while their correspondences transformed from the reference (image in the left of each sub figure) to the second (image in the right of each sub figure) using ground truth are in blue. The region correspondences computed based on ground truths and overlap errors are highlighted with blue lines whereas matches identified as correct using descriptors are highlighted with green lines. For the purpose of clarity, only few correspondences are shown.

5.3.9.1 Discriminative power

The evaluation results of the discriminative power under JPEG compression are depicted in Figs. 5.73, 5.74, 5.75, and 5.76.

The scores are obtained for feature matching between the reference and 6th images. This latter is shown in the right side of Fig. 5.72. It corresponds to the most degraded image within the images of `ubc`'s scene.

The figures show all curves obtained with the `nearest-neighbor` (*i.e.*, Figs. 5.73a, 5.74a, 5.75a, and 5.76a) are mostly situated toward the upper left-hand side of the ROC space. This means the discriminative power of descriptors is much better under JPEG compression than under other deformations (*e.g.*, rotation, scale change, image blur, etc).

On the other hand and in contrast with image blur, we observe that the discriminative power of SPIN and CC is noticeably ameliorated when the SSC information is added, which is not the case for SIFT as we can see across comparing the SIFT and SIFT-Based-SSC curves.

Despite this, we find SIFT-Based-SSC still being in the top spot over all evaluations, and besides, the ranking of CC-Based-SSC is much better than that of CC, yet more it wins SC and GLOH in the case of Fig. 5.74a where it is placed 3rd behind SIFT.

We also notice that the role of the SSC component is more effective with CC than with SPIN. This can be seen, for example, in Fig. 5.75b, in which while the SPIN and CC curves are almost aligned, the curve of CC-Based-SSC goes much far above that of SPIN-Based-SSC.

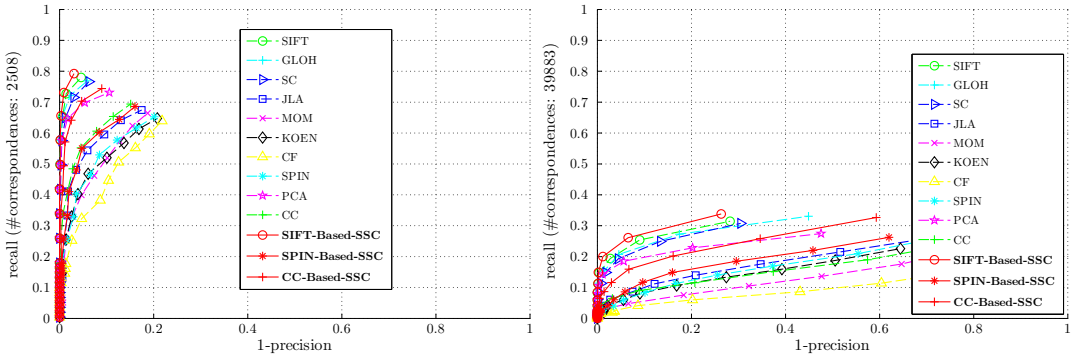
These differences seem more important when examining the gains in the number of correct matches obtained separately for each SSC-based descriptor as highlighted in Tab. 5.9.

In this table, the number of correct matches obtained with the local components, *i.e.*, SIFT, SPIN, and CC, are reported along the first row while the percentages of the number of correct matches gained when adding the semantic-context components are recorded along the second row. Furthermore, we show the ranking changes along the third row. The scores are established for number of correspondences of 39932 and with respect to a precision of 0.2.

This table reflects how well the CC descriptor is positively influenced by the SSC information where the improvements manifest clearly to be much better than those of SPIN-Based-SSC and SIFT-Based-SSC descriptors.

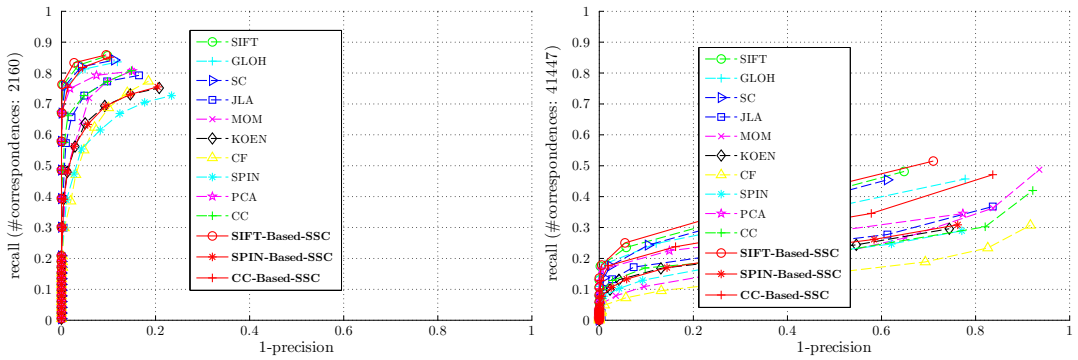
TAB. 5.9: Gains in the number of correct matches when adding semantic-context informations in the case of JPEG compression (scene of *ubc*). The number of correct matches obtained with the local components (SIFT, SPIN and CC) are reported along the first row while percentages of number of correct matches gained when adding semantic-context components are recorded along the second row. The ranking changes are exhibited along the third row. The scores are established for number of correspondences of 39932 and with respect to a precision of 0.2. The descriptors are computed for Hessian-Affine regions and then matched with the threshold-based matching strategy.

	SIFT-Based-SSC	SPIN-Based-SSC	CC-Based-SSC
#Correct matches	**11980**	6590	6588
Gain ↑ (%)	+03%	+09%	+**63**%
Ranking change	$1 \rightarrow 1$	$7 \rightarrow 6$	$\mathbf{8} \rightarrow \mathbf{4}$



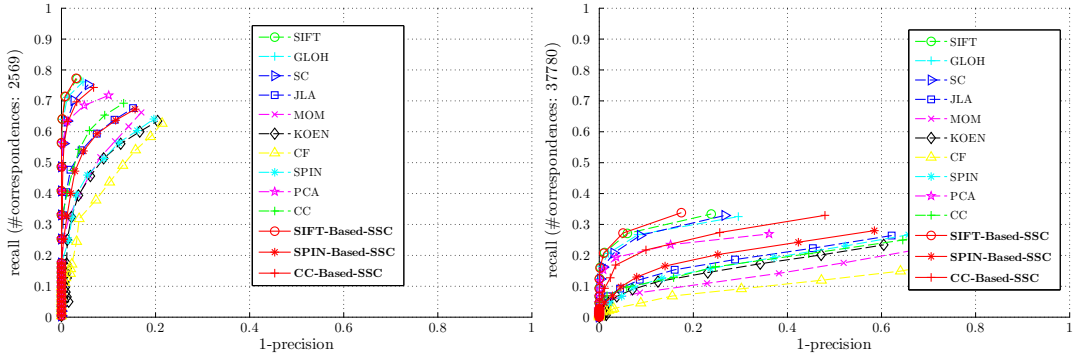
(a) (b)

FIG. 5.73: Evaluation results of the discriminative power under JPEG compression. The results are obtained for the structured scene of *ubc* of Fig. 5.2h. The descriptors are computed for Harris-Laplace regions and matched using (a) the nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2508 and (b) 39883 correspondences.



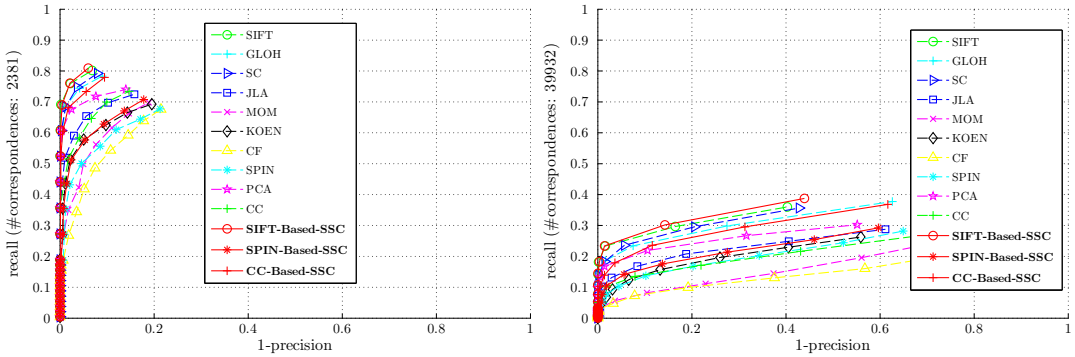
(a) (b)

FIG. 5.74: Evaluation results of the discriminative power under JPEG compression. The results are shown for the structured scene of *ubc* of Fig. 5.2h. The descriptors are computed for Hessian-Laplace regions and matched using the (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are with respect to (a) 2160 and (b) 41447 correspondences.



(a) (b)

FIG. 5.75: Evaluation results of the discriminative power under JPEG compression. The results are shown for the structured scene of *ubc* of Fig. 5.2h. The descriptors are computed for Harris-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2569 and (b) 37780 correspondences.



(a) (b)

FIG. 5.76: Evaluation results of the discriminative power under JPEG compression. The results are shown for the structured scene of *ubc* of Fig. 5.2h. The descriptors are computed for Hessian-Affine regions and matched using (a) nearest-neighbor and (b) threshold-based matching techniques. The recall scores are computed with respect to (a) 2381 and (b) 39932 correspondences.

5.3.10 Computation Time

For the computational time (in terms of wall-clock time), we reported 1.9 ms/per-feature for computing the SSC component of SIFT-Based-SSC. This is approximately 41% of SIFT computational time (4.8 ms/per-feature). The recorded times are obtained on the highly textured scene of `zeriba`.

This is very motivating, since the focus here is more upon the performance than efficiency.

5.4 Conclusion

In this chapter, the concept of SSC is used to address the problem of 2D image feature matching on real data sets composed of different types of images, subjected to different geometric deformations and imaging conditions.

The reported results showed SSC-based descriptors to perform significantly above the expected performance. It clearly illustrated the effectiveness of the approach, under particular conditions, to turn impractical descriptors into well suited descriptors. This has been demonstrated for different geometric image transformations.

In addition, it is also noticed the usefulness of the SSC approach under imaging condition changes is less better but still outperforms other descriptors as we showed for SIFT-Based-SSC under image blur, illumination change, and JPEG compression.

Furthermore, we observed how well the performance of a basic descriptor like cross-correlation (CC) is highly increased when adding the semantic-context component as in cases of geometric transformations and image blur.

Chapter 6

Experimental Results – 3D Domain: Matching and Alignment of Multiple Range Images

6.1 Introduction

In this Chapter we test the effectiveness of the proposed SSC feature descriptor by addressing the problem of *multiview surface* matching. As customary [Huber 03], registration[1] is carried out pairwise first, and then a simple strategy for treating multiple range images[2] is implemented.

6.2 Pairwise Registration

We propose our feature matching strategy to solve robustly the issue of pairwise 3D view pre-alignment.

Given a pair of 3D views, feature matching estimation is carried out by comparing the descriptors of each feature point of the first view (range), with all the feature points of the second range. In this fashion a graph of point-to-point similarities is built and the correspondences are estimated based on bipartite graph matching concept [Duda 01] by using Hungarian algorithm [Frank 05]. This algorithm is applied on the global similarity matrix, C, defined in Equation 3.4.

Then, in order to remove false matches, the standard RANSAC [Fischler 81] algorithm is implemented. It imposes the rigid constraint among two views. The output of RANSAC is obtained as pre-final feature-point matching, *i.e.*, the pre-alignment. Based on this pre-alignment the ICP refinement algorithm is then applied to obtain the pairwise view alignment.

In addition, to allow the SSC component to be influenced from the whole 3D view to a small neighborhood, the weighting function, $g(\cdot)$ of Equation 3.6 is defined as Gaussian kernel. This approach is especially useful in the context of partial 3D view matching since furthest points are likely to be occluded.

Since the Gaussian function relies heavily on the scale parameter, σ, we implement a greedy approach evaluating a set of values $\sigma_s \in I$ to determine the best scale σ_{best}. This is selected as that recording the minimal pairwise registration error[3].

[1] In this paper, the terms *alignment* and *registration* are used equivalently.

[2] Here, a view refers to a partial 3D view, also called range image.

[3] The registration error is computed by summing the residual errors of all corresponding points between the two views after the pre-alignment, where the correspondences are computed by closest point.

6.3 Treating Multiple Range Images

To extend the pairwise registration to multiple registration, we build the *registration matrix* $M_{x,y}$. This matrix stores the *registration score* for each pairwise view matching *i.e.*, the pre-alignment error between range images R_x and R_y.

We also suggest a strategy for estimating the best pair of views to be registered. We binarize the matrix $M_{x,y}$ by using a threshold obtained experimentally. Then we estimate the best path which connects all the range and minimize the sum of registration scores (*i.e.*, similar to [Novatnack 08] by maximum spanning tree).

Finally, the registrations of the selected pairs are refined by ICP [Besl 92] and all the views are put to the global reference system by simply concatenating subsequent pairwise rigid transformations.

6.4 Experimental Setup

The performance of 3D SSC-based descriptor is evaluated for matching and alignment of multiple range images by performing three different experiments. These include, pairwise matching and registration of range images using `(i)` one-information-based SSC and `(ii)` multiple-information-based SSC descriptors, and `(iii)` an automatic multiple range images registrations.

The experiments are conducted based on the following setting:

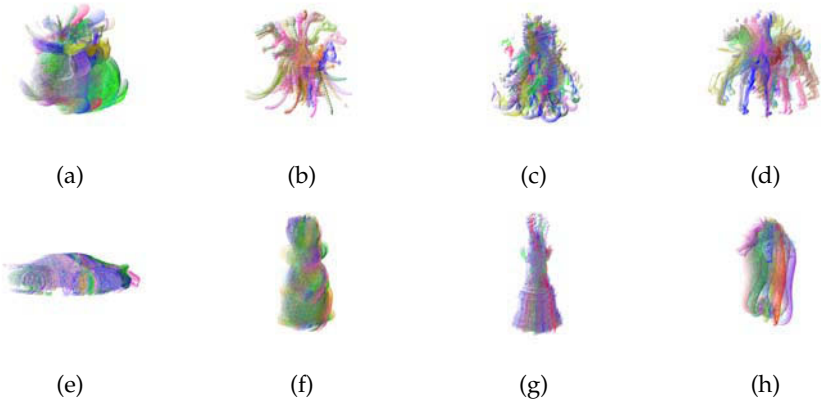
6.4.1 Data Set

The evaluations are carried out using range images collected from the standard database of Stuttgart university available on-line[4]. These contain 10 full models constituting 120 range images in total, *i.e.*, each model is composed of 16 range images obtained from different azimuthal viewpoint angles, as illustrated in Figs. 6.1.

The range images are produced under uncontrolled conditions using the 3D scanners, Cyberware 3030 MS and XYZ RGB. Under such conditions which reflect a real environment of scanning process, several errors can appear. The errors related to outliers, scan misalignments, and device as well as systematic errors are few examples.

These errors can cause many approaches to be ineffective. Through exploiting of the proposed SSC approach, we expect better performance can be achieved.

[4]http://range.informatik.uni-stuttgart.de/htdocs/html/



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

(e) (f) (g) (h)

FIG. 6.1: Examples of range images used to evaluate the 3D SSC-based descriptor. Each model in the data set is composed of 16 range images obtained from different azimuthal viewpoint angles. We show here those for (a) Bunny, (b) Dino, (c) Dragon, (e) Porsche, (f) Hasi, and (g) Liberty.

6.4.2 Evaluation Criteria

We use three registration measures, inlier percentage (*i.e.*, correct matches percentage), pre-alignment error, and alignment error.

The inlier percentage is computed as the ratio of the number of inliers (*i.e.*, obtained by RANSAC algorithm) to the number of putative matches (*i.e.*, the potential matches obtained by the Hungarian algorithm).

The pre-alignment error computed as *RANSAC Root Mean Squared Error*, abbreviated as RANSAC-RMSE. This measures the average of error squares occurring after applying RANSAC to align two range images.

The alignment error, ICP-RMSE, is similar to the latter whereas it is computed after applying ICP final alignment (*i.e.*, refinement).

6.4.3 Descriptors

This proposed SSC descriptor is compared to the both standard 3D Shape-Context [Frome 04] and Spin-Image [Johnson 99] descriptors, early presented in Section **Related Work**.

Briefly, a 3D Shape-Context descriptor is computed as a histogram encoding the point distribution around a reference feature point, as function of three quantized spherical coordinates: radial distance, r, elevation angle, θ, and azimuthal angle, ϕ. In our

evaluations, we use an histogram of size $10 \times 10 \times 10$.

For Spin-Image, the descriptor is also computed as a histogram, but with respect to two parameters only. Thus, each histogram is a function of the quantized cylindrical coordinates α and β, which are the positive radial and signed elevation coordinates, respectively. The experiments use a histogram of 10×10.

6.4.4 Computation of Local Descriptors and Context Components

For shape-index measure (information), the L is a histogram of size 6×8, whereas that related to β-value is data dependencies. That is to say, the histogram size of L component depends on mesh resolution of the range image being treated. For instance, we computed a histogram of size 215×408, for *Bunny*'model.

The SSC component is of size $6 \times K$, in which the number of visual words, K, is obtained as an estimate of the average number of features per range image, *e.g.*, we obtained $K = 50$ for *Bunny*'s model.

6.4.5 Matching and Registration Strategies

We adopt range image feature matching based on bipartite graph concept, for which the Hungarian algorithm is applied. This requires a similarity matrix (*i.e.*, Equation 3.4) to be previously computed, *i.e.*, as an input of the algorithm.

To reject false matches, the RANSAC alignment algorithm is then applied by imposing a rigid body transformation between each pairwise range images. Thus, the final set of correct matches (inliers) are provided. This is used later by the ICP refinement to produce the final pairwise range images alignment.

6.5 Results and Discussion

As mentioned above, the performance of 3D SSC approach is evaluated within three experiments. The first two experiments are related to pairwise view registrations.

In the first experiment of Section 6.5.1, we use the shape-index measure, si, to build the local descriptor, L, and the global similarity matrix, C, is given as follows:

$$C(p_1, p_2) = C_L^{si}(p_1, p_2) + C_{SSC}^{si}(p_1, p_2). \tag{6.1}$$

In the second experiment of Section 6.5.2, The global descriptor is computed around two different geometric measures, shape-index, si, and β-value, bv. That is, G is a combination of four different components, for which the matrix, C, is written as:

$$\begin{aligned} C(p_1,p_2) = C_L^{si}(p_1,p_2) + C_L^{bv}(p_1,p_2) + \\ C_{SSC}^{si}(p_1,p_2) + C_{SSC}^{bv}(p_1,p_2). \end{aligned} \tag{6.2}$$

The weighting coefficient, w, appearing in Equation 3.4 is discard, *i.e.*, the different components are contributing equally. The G descriptor constructed thus is then compared to the standard approaches of 3D Shape-Context and Spin-Images.

In the third experiment of Section 6.5.3, the SSC-based descriptor tested during the second experiment (*i.e.*, Equation 6.2) is used to perform an automatic multiple range images registration, and its performance is compared to those obtained with Shape-Context and Spin-Images.

Before going into details, Fig. 6.2 shows an example of feature matching between two range images, in which the putative matches (after applying Hungarian algorithm) are highlighted with blue lines while the inliers (after applying RANSAC algorithm) are shown with red lines.

During the experiments, the scale, σ_s related to the Gaussian function, $g(\cdot)$, used to compute the SSC component is automatically adjusted, *i.e.*,

$$\sigma_s \in \{0.1,\ 0.6,\ 1.1,\ 1.6, \ldots, 4.6\}, \tag{6.3}$$

while the other parameters remain fixed. It appears that the only revealing parameters are the scale, σ_s, and number of visual words, K. This is set as the average number of feature points per range image.

6.5.1 One-Information-Based SSC Descriptor

In this evaluation, the performance of the SSC-based descriptor, G, obtained around the matrix distance of Equation 6.1, is evaluated and compared to those of L and SSC components.

The obtained results are depicted in Figs. 6.3, 6.5, and 6.5. Following are some observations reported with respect to the registration measures: percentage of inliers, pre-alignement error, and alignment error.

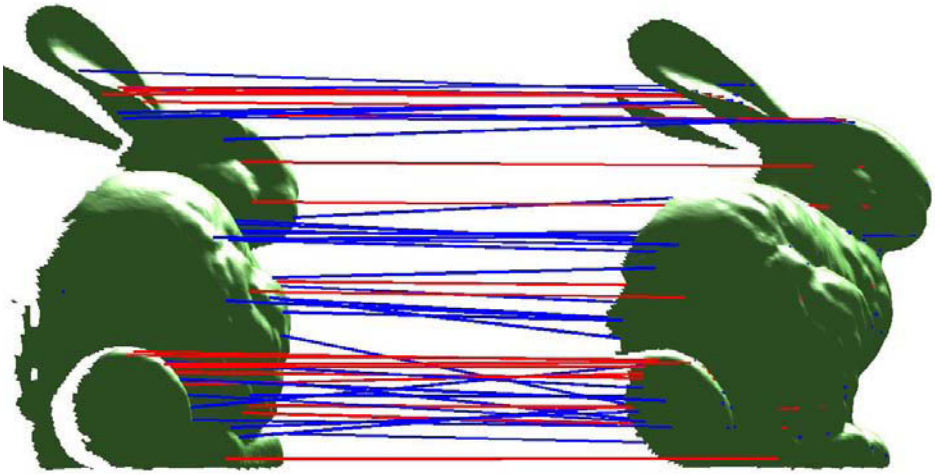


FIG. 6.2: An example of feature matching between two range images, in which the putative matches (after applying Hungarian algorithm) are highlighted with blue lines while the inliers (after applying RANSAC algorithm) are shown with red lines. Here, the range images are displayed as range surfaces, *i.e.*, after performing triangulation and shaded rendering.

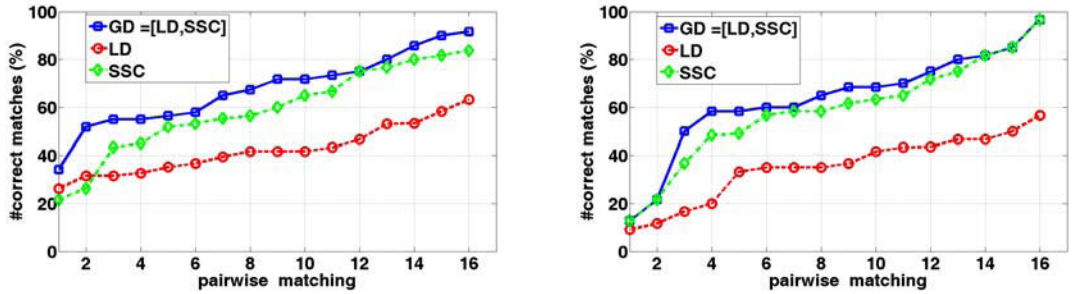
6.5.1.1 Inlier percentages

In Fig. 6.3, the considerable gap in the percentage of correct matches looks reasonably straightforward, where the curves of G is always above those of L and SSC components. It is easy to observe that SSC component performs better than L.

This is quite expected, since the contextual signature, SSC, is often well-suited for range images, which are composed of point cloud with local discontinuity structures.

The aforementioned illustrates that G is more discriminative, and thus the number of accurate putative matches obtained after applying Hungarian algorithm is high enough for RANSAC to produce a higher number of inliers as well.

We note that similar results are recorded on the other models and the performance of G on high mesh resolution models are observed to be better than those on models of less mesh resolutions. For instance, G performs better on Bunny than on Dino as shown in Fig. 6.3a and Fig. 6.3b.



(a) Bunny (b) Dino

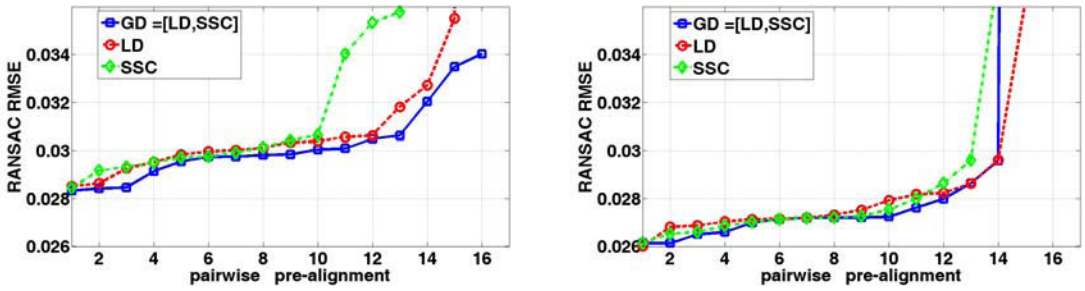
FIG. 6.3: Evaluation of range-image feature matching according to the percentage of inliers (correct matches). The results are obtained for (a) *Bunny* and (b) *Dino* models. On the x-axis the pairwise range image pre-alignments are sorted in ascending order from 1 to 15 since the model contains 16 range image

6.5.1.2 Pre-alignment error (RANSAC-RMSE)

The results related to pre-alignment errors are displayed in Fig. 6.4. These are obtained on two models with different mesh resolutions. We observe that the pre-alignement error computed for G is less than those for L and SSC.

Much as the previous evaluation (w.r.t. inlier percentages), G produces more accurate pre-alignemets on Bunny than on Dino as illustrated in Fig. 6.4a and Fig. 6.4b, respectively. Since the resulting pre-alignement based on G is always more accurate, thus it is

more easy to be directly refined by any ICP-based algorithm.

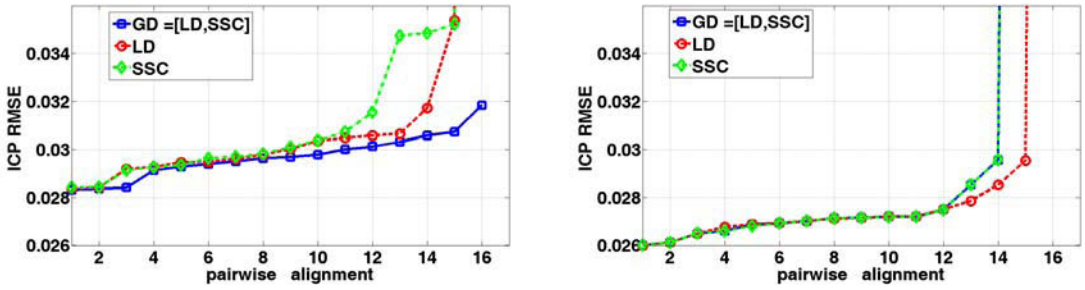


(a) Bunny (b) Dino

FIG. 6.4: Evaluation of pairwise range image pre-alignment according to RANSAC-RMSE. The results are obtained for (a) *Bunny* and (b) *Dino* models. On the x-axis the pairwise range image pre-alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images

6.5.1.3 Alignment error (ICP-RMSE)

The evaluation results obtained according to the final alignment error are demonstrated in Fig. 6.5. We reported the final alignment error after applying ICP refinement less important with G than with L and SSC taken separately. This is because of good pre-alignments (RANSAC alignments) provided after combined L and SSC components. Similar to the pre-alignment evaluation, the final alignment is noticed to be in general better on models of high resolution meshes than those of less resolution meshes. Fig. 6.5a and Fig. 6.5b are examples.



(a) Bunny (b) Dino

FIG. 6.5: Evaluation of pairwise range image alignment according to ICP-RMSE. The results are obtained for (a) *Bunny* and (b) *Dino* models. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images

6.5.2 Multiple-Information-Based SSC Descriptor

In this evaluation, G descriptor computed on both shape-index and β-value geometric measures (as given in Equation 6.2), is compared to 3D Shape-Context and Spin-Images descriptors.

The descriptors are evaluated according to the RANSAC and ICP alignment errors defined early. Our evaluation proceeds by computing the registration matrix, $M_{x,y}$, that contains the alignment error for each pair of images (we used 16 range images of each model).

By varying the scale, σ_s, (weighting the spatial extent of SSC component) as given in the statement 6.3, we select the best, σ_{best}, for which the RANSAC pre-alignment error is the minimum.

Then we record the corresponding RANSAC and ICP alignment errors as well as the associated transformation. We find the best path of a weighted graph (*i.e.*, the registration matrix, $M_{x,y}$) by using the *maximum spanning tree* algorithm [Wu 04].

The test models are intentionally selected to obtain different type of range images with diverse properties, *i.e.*, contain various type of features, surfaces (*e.g.*, planar), symmetries, and curvatures. The evaluation results obtained on eight of these models are depicted within Figs 6.6 . . . 6.13.

The obtained results illustrates without doubt the best performance of our descriptor compared to those of 3D Shape-Context and Spin-Image. This holds on different evaluated models when the approximate pairwise registration (*i.e.*, RANSAC pre-alignment) obtained with our descriptor still the best for the almost pre-alignments.

Moreover, in many cases the 3D Shape-Context and Spin-Images failed completely to produce an accurate enough pre-alignment, which is required by any ICP-based alignment algorithm in order to converge.

Fig. 6.11 is an example demonstrating a number of pairwise alignment based on 3D Shape-Context and Spin-Images performed unsuccessfully while our descriptor provides accurate pre-alignments in all cases.

In order to demonstrate the robustness of our descriptor according to the overlap area, we computed the best scale, σ_{best}, for each pairwise alignment. The corresponding alignment error is reported. We use *large*, *medium*, and *small* to define the amount of overlap, which is proportionally estimated from σ_{best} (*i.e.*, the larger σ_{best} the larger the overlap).

The computed pre-alignment error for each pairwise range image is compared to

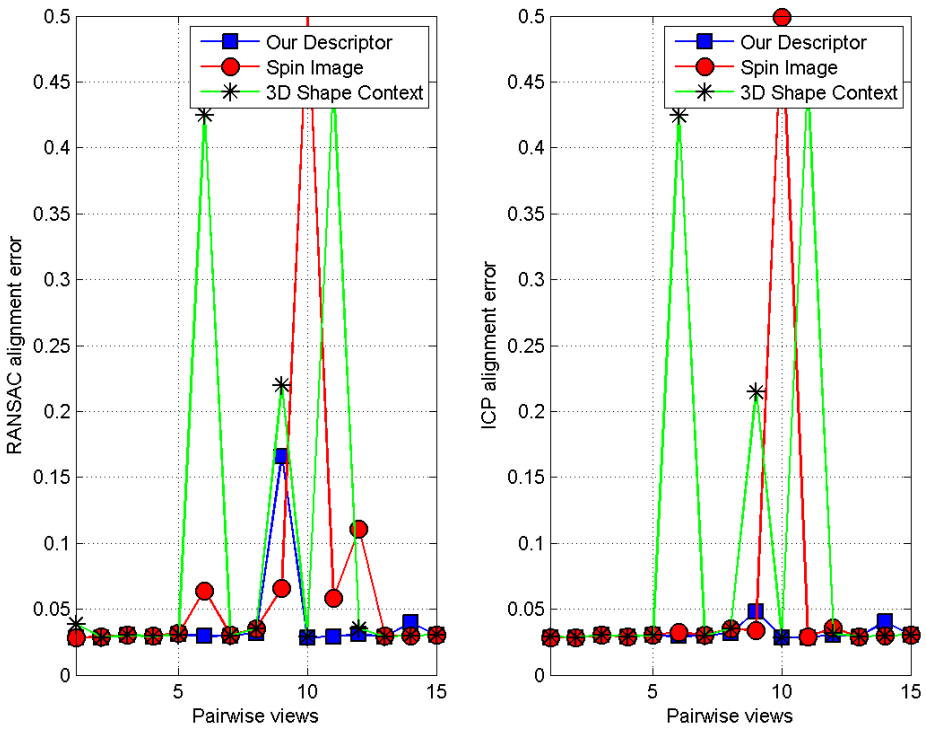


FIG. 6.6: Evaluation of pairwise alignment errors for *Bunny*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

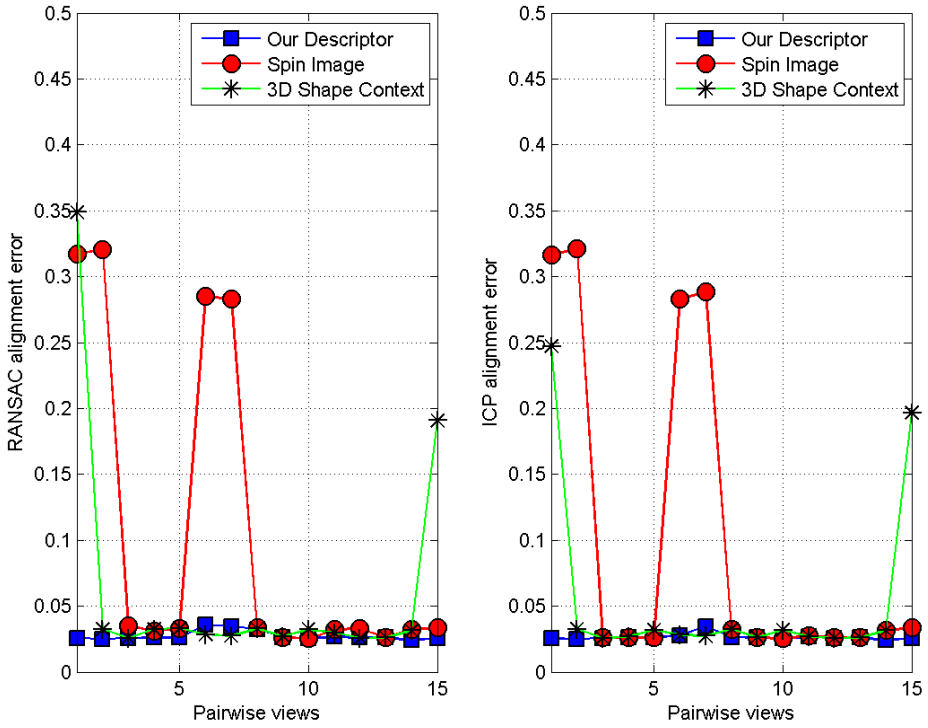


FIG. 6.7: Evaluation of pairwise alignment errors for *Bull*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

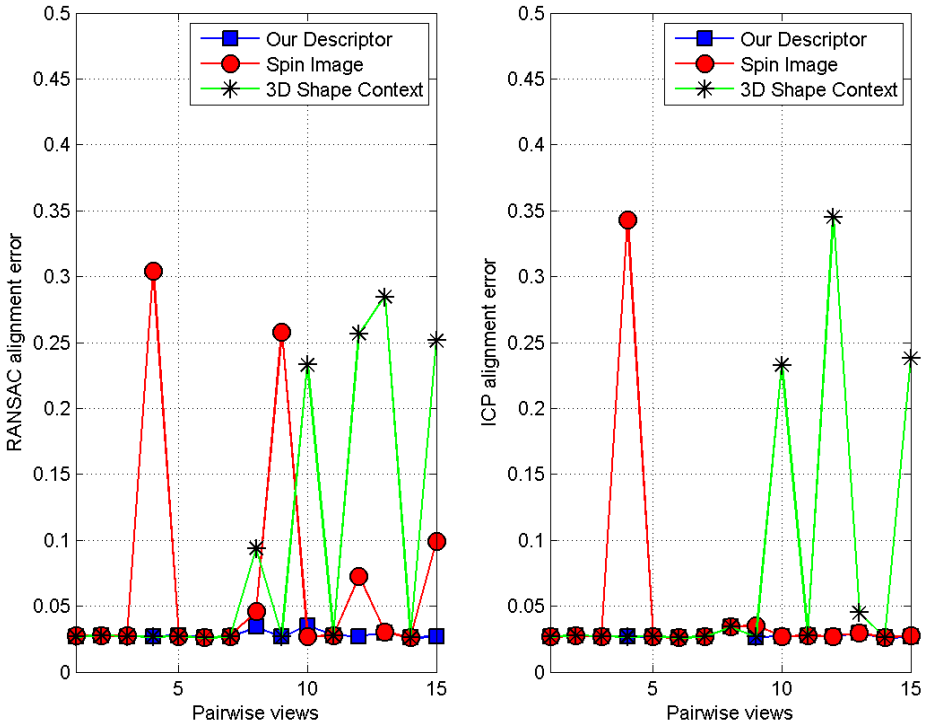


FIG. 6.8: Evaluation of pairwise alignment errors for *Dino*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

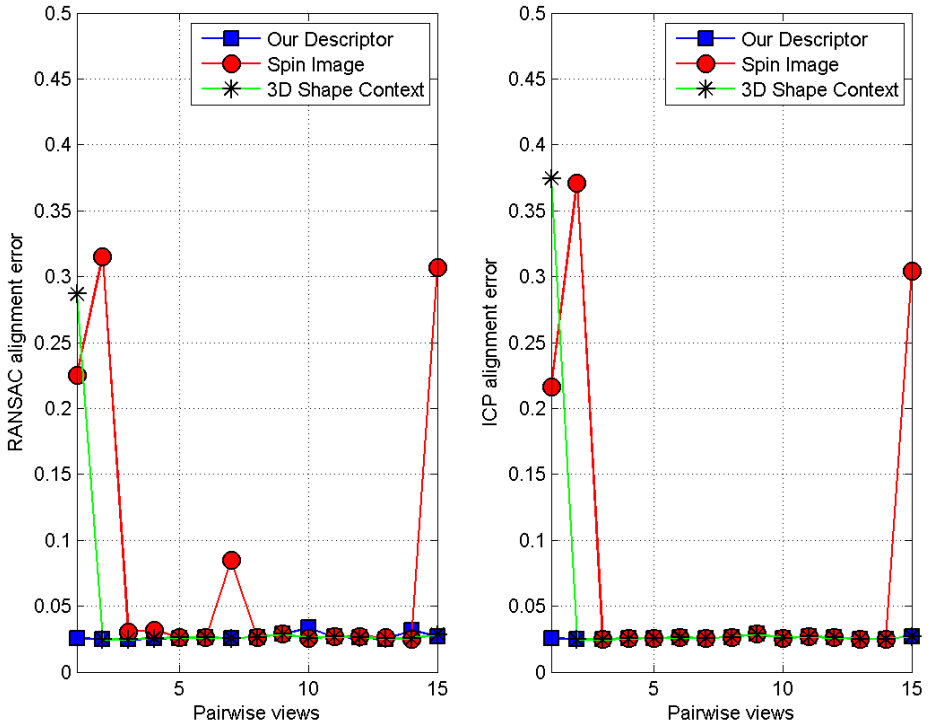


FIG. 6.9: Evaluation of pairwise alignment errors for *Dragon*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

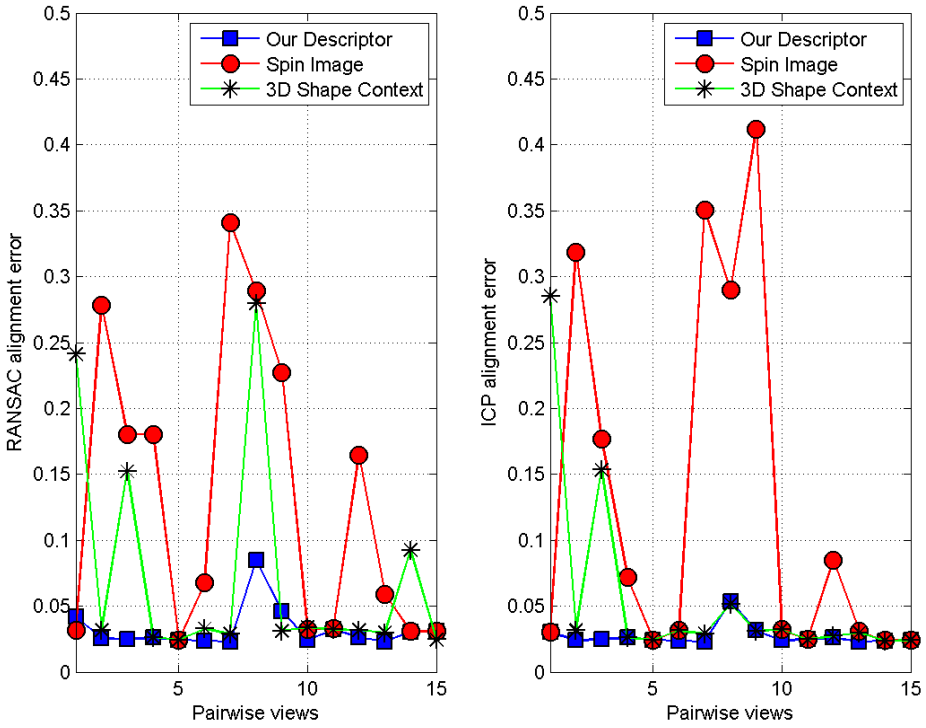


FIG. 6.10: Evaluation of pairwise alignment errors for *Female*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

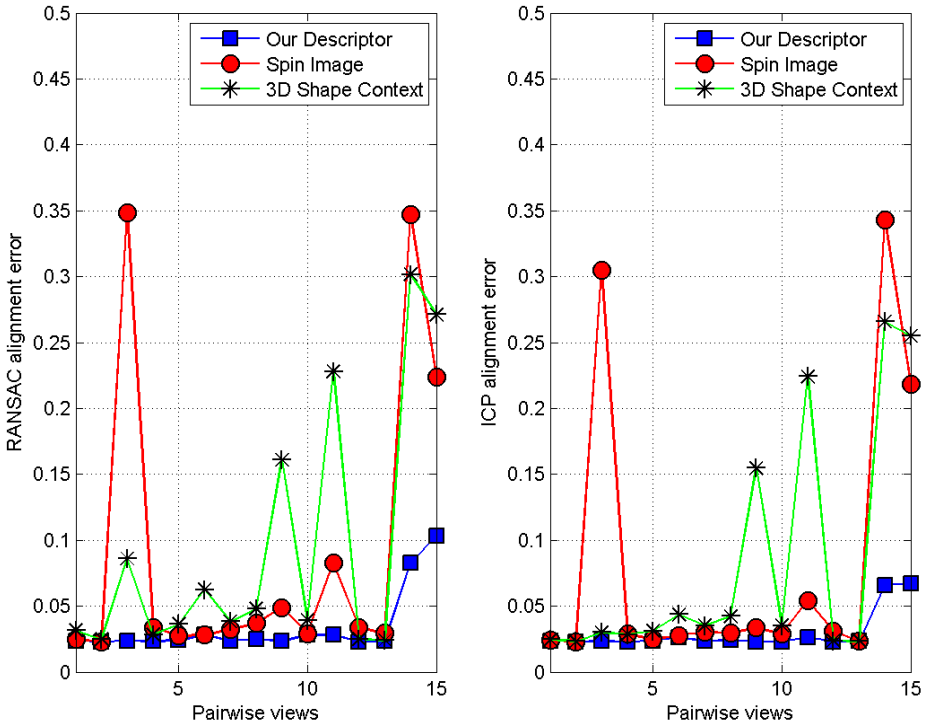


FIG. 6.11: Evaluation of pairwise alignment errors for *Hasi*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

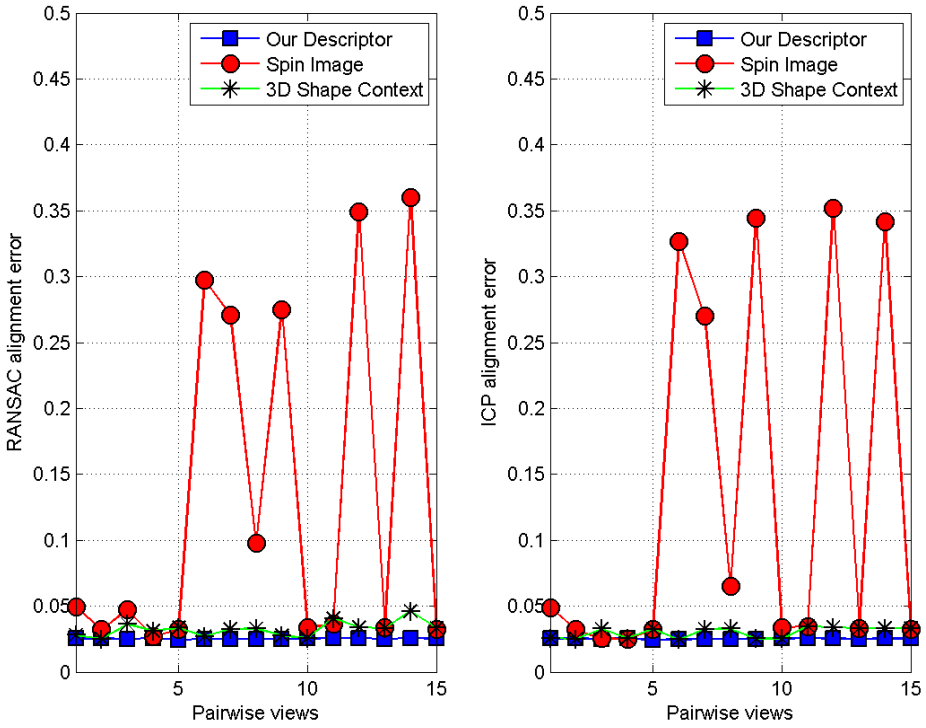


FIG. 6.12: Evaluation of pairwise alignment errors for *Screwdriver*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

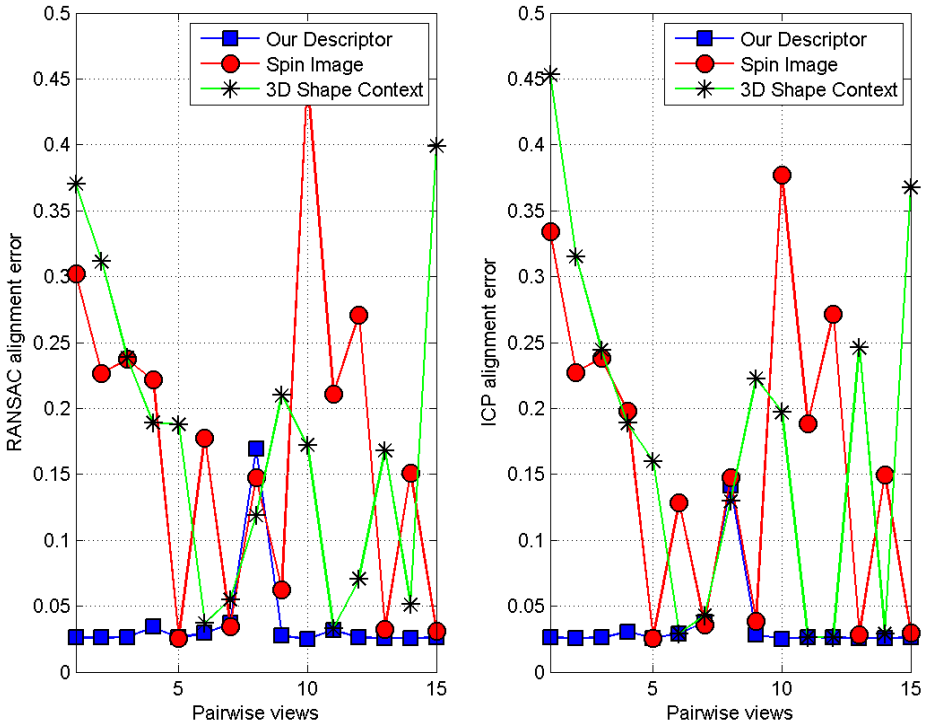


FIG. 6.13: Evaluation of pairwise alignment errors for *Seahorse*'s model. The errors are related to (*Left*) RANSAC pre-alignment and (*Right*) ICP alignment. On the x-axis the pairwise range image alignments are sorted in ascending order from 1 to 15 since the model contains 16 range images.

those obtained with 3D Shape-Context and Spin-Image. Tabs. 6.1 . . . 6.4 summarize the results.

For each model we selected 4 different pairwise alignments according to the σ_{best} value (*i.e.*, large, medium, or small overlap). Note that the 3D Shape-Context and Spin-Images approaches provide more accurate alignments for large overlaps only, while for the small to medium overlaps, our descriptor seems to be significantly better.

Thus, we observe that for low (*i.e.*, small and medium) overlaps, the pre-alignment errors reported with our descriptor are mostly smaller than those of other descriptors. This is clearly illustrated, for example, in Tabs. 6.2 and 6.4.

Even though it records slightly higher errors than 3D Shape-Context and Spin-Image for large overlaps (*e.g.*, Tab. 6.4), it is still operational to provide enough accurate pre-alignments [5] necessaries for ICP-based refinement to be directly and successfully performed.

This is not the case with the other descriptors for many pairwise pre-alignments where the pre-alignement errors provided by 3D Shape-Context and Spin-Image are above the specific threshold. [5]

TAB. 6.1: Evaluation of our SSC-based descriptor (global) robustness for Bunny's model according to the overlap area.

			RANSAC alignment error		
Pairwise alignment	σ_{best}	overlap	**Our Descriptor**	**Spin Image**	**3D Shape Context**
08 ← 07	3.1	large	0.0310	**0.0283**	0.0388
14 ← 12	1.6	medium	**0.0292**	0.0583	0.4479
03 ← 02	0.6	small	**0.0301**	0.0637	0.4248
09 ← 10	4.6	large	**0.0284**	0.0306	0.0304

TAB. 6.2: Evaluation our SSC-based descriptor (global) robustness for Bull's model according to the overlap area.

			RANSAC alignment error		
Pairwise alignment	σ_{best}	overlap	**Our Descriptor**	**Spin Image**	**3D Shape Context**
07 ← 08	0.1	small	**0.0256**	0.3168	0.3489
11 ← 12	4.1	large	**0.0260**	0.0329	0.0333
16 ← 02	1.6	medium	**0.0267**	0.0323	0.0299
09 ← 10	3.6	large	**0.0256**	0.0348	**0.0256**

The main reasons behind the best performance of our descriptor are:

[5] We found experimentally that for pre-alignment errors below 0.06, ICP-based algorithm converges to the good solution.

TAB. 6.3: Evaluation of our SSC-based descriptor (global) robustness for Dino's model according to the overlap area.

			RANSAC alignment error		
Pairwise alignment	σ_{best}	overlap	**Our Descriptor**	**Spin Image**	**3D Shape Context**
16 ← 02	0.1	small	0.0269	0.0727	**0.2568**
03 ← 04	3.6	large	0.0271	**0.0260**	**0.0260**
02 ← 03	1.6	medium	**0.0275**	0.0301	0.2847
07 ← 08	4.6	large	**0.0261**	0.0276	0.0277

TAB. 6.4: Evaluation of our SSC-based descriptor (global) robustness for Dragon's model according to the overlap area.

			RANSAC alignment error		
Pairwise alignment	σ_{best}	overlap	**Our Descriptor**	**Spin Image**	**3D Shape Context**
04 ← 05	4.1	large	0.0263	0.0263	0.0263
13 ← 14	2.1	medium	**0.0258**	0.2252	0.2869
14 ← 15	1.1	small	**0.0246**	0.3149	0.0247
03 ← 04	2.6	medium	**0.0258**	0.0846	**0.0258**

- Combining local descriptor with context component can resolve the ambiguities that may appear in each other (*e.g.*, locally where a view has similar surface elements).
- By using multiple sources of information, *e.g.*, shape-index and β-value, we are able to compensate for apparent (or intrinsic) defects of one component by retrieving information from another (*i.e.*, from local to context components and vice versa).
- Another strong point in our approach is related to rotation invariance of our descriptor. It means that we can use a simplified histogram-like descriptor. Therefore, we avoid a high loss of information [Malassiotis 07] (*i.e.*, going from a 3D to a 2D representation) and gain more in robustness.

These mentioned arguments may explain the superior performance of our approach in the case of small overlap illustrated in the above tables.

6.5.3 An Automatic Registration of Multiple Range Images

Based on the same global descriptor, G, previously computed (*i.e.*, multiple-information-based SSC of Section 6.5.2), we conducted an automatic multiple range image registration. We evaluated a collection of 10 models, consisting of 16 images in each model.

The images are chosen in such a way that ICP would diverge by requiring a robust pre-aligning strategy.

Figure 6.14 shows the results of 4 models. Note that the registered models are further refined with an ICP-based algorithm. It can be seen across the 2nd and 3rd columns of each row that these registered models do not have noticeable defects.

This because of the approximate alignment (*i.e.*, with the RANSAC-based algorithm) which is obtained is very accurate, and thus can be directly and successfully refined with an ICP-based algorithm.[6]

Finally, similar automatic multiple range image registration are performed using 3D Shape-Context and Spin-Image. The results are illustrated in Tab. 6.5, in which the three approaches are compared according to the percentages of the number of pairwise alignments performed successfully (*i.e.*, ICP-algorithm converges towards the good solution) for each tested model.

TAB. 6.5: Evaluation of automatic multiple range image registration according to the percentages of the number of pairwise alignments, which are performed successfully *i.e.*, ICP-algorithm converges towards the good solution.

Model	**3D Shape-Context**	**Spin-Image**	**Our Descriptor**
Bunny	81.2	87.5	**100**
Dino	75.0	75.0	**100**
Bull	93.7	87.5	**100**
Dragon	**100**	**100**	**100**
Liberty	68.7	93.7	**100**
Hasi	**100**	93.7	**100**
Porshe	81.2	**100**	93.7
Mole	93.7	87.5	**100**
Copter	**100**	81.2	87.5
Screwdriver	81.2	68.7	**93.7**
Average	87.4	87.4	**97.4**

This table demonstrates that out of 10 tested models, we are able to align (*i.e.*, overlay) successfully [7] the range images of 7 models when using our descriptor. Whereas we report only 3 and 2 models when adopting 3D Shape-Context and Spin-Image approaches, respectively.

Furthermore, the registered (*i.e.*, aligned) models obtained with our approach appear to be very accurate, without visible imperfections, as shown in Fig. 6.14.

[6] An ICP-based algorithm requires a sufficient overlap among the views and a coarse pre-registration to avoid getting stuck into a local minimum.

[7] Those obtain percentages of 100%, which means the 16 range images composing the model are aligned successfully.

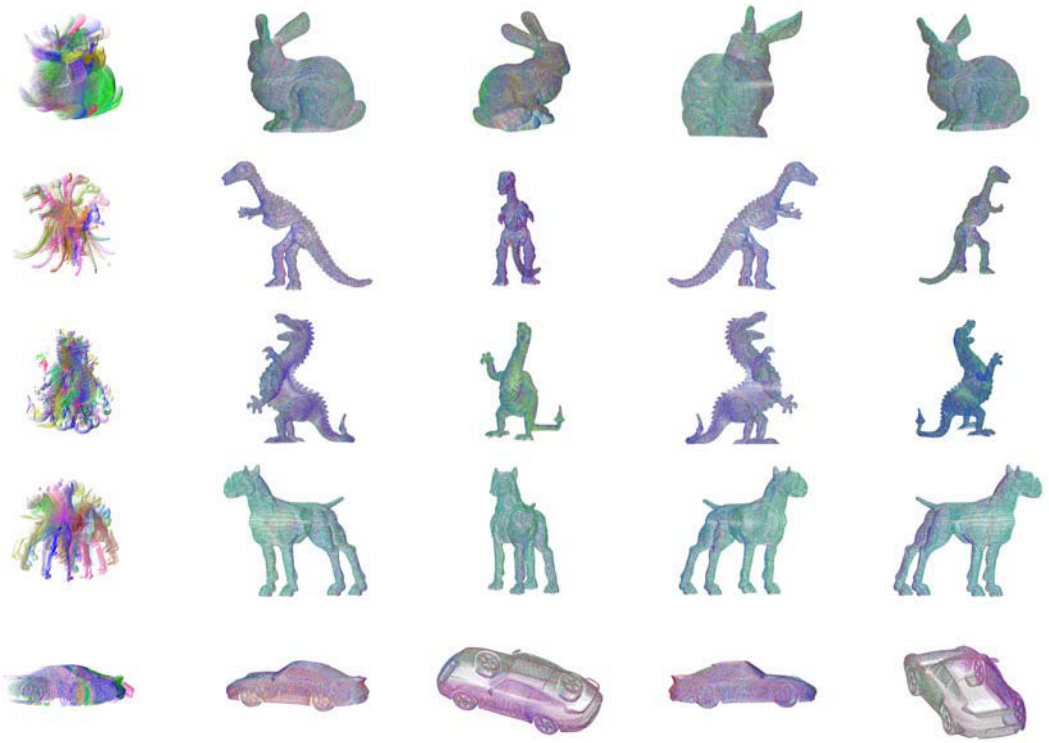


FIG. 6.14: An automatic multiple range images registration based on the SSC descriptor. Here, along the 1st *column*, four sets of range images before registration are displayed, while along the 2nd to 5rd columns of each row, the registered range images seen from 4 different angles. These registered models are very accurate; indeed, they appear without visible defects.

6.6 Conclusion

In this chapter, we showed that the SSC is a successful yet simple concept, to implement a powerful feature descriptor which resolves efficiently the problem of pre-alignemnt, *i.e.*, approximate alignment by postapplication of a RANSAC-based method. This pre-alignemnt becomes accurate enough to be directly refined by an ICP-based registration method. We also introduced an approach to automatically estimate the overlap area. We used the scale for determining which views overlap. Based on the best scale, we find the best pairs of views to be registered to each other for fully automatic registration. Our experiments demonstrated the efficiency of our descriptor for the case of a small overlap.

Chapter 7

Conclusion

7.1 Introduction

We have proposed across this manuscript, an extension of feature description and matching strategies, through presenting an original approach to *learn* the meaning of local features, which leads in turn to estimate the *semantic* of the local features. This semantic is then exploited, in conjunction with the *bag-of-words* paradigm, to build a powerful feature descriptor.

The approach consists, first, in extracting the semantic information around local information collected on different images. Thus, the local descriptors computed on different images are grouped and then clustered to generate what we call, semantic features (or visual words) – in fact, these are the different obtained clusters.

Inspired by bag of words paradigm (BoW), these semantic features are then accumulated inside spatial concentric shells to obtain an histogram-based representation of Semantic-Shape-Context component. This, is next concatenated to the local component to obtain finally our proposed SSC-based feature descriptor.

The effectiveness of the SSC concept is illustrated for two different real machine vision applications. The first is a 2D-domain problem which addresses the problem image features matching. The second is a 3D-domain issue, which involves matching and alignment of multiple range images.

The experimental results showed our approach performs much better compared to other methods. For both the 2D-domain and 3D-domain addressed problems, the higher performance scores are recorded by our proposed SSC-based descriptors. This is specially illustrated on the images extracted under hard and uncontrollable conditions, where the standard descriptors quickly achieve their limitations and turn out to be unworkable in some cases.

The remainder of this chapter summarizes the content of our contribution regarding the usefulness of SSC concept in both the 2D and 3D domains.

7.2 SSC Approach in 2D-Domain

Our evaluations in 2D-domain have been conducted to resolve an important and crucial problem, omnipresent in many computer vision applications. This consists in matching features on images, subjected to real geometric transformations and relevant imaging conditions.

The evaluations are previously designed to compare the performance of three variant

of SSC-based descriptors to those obtained with ten of well-known and most usable state-of-the-art descriptors. The performances are evaluated with respect to both the discriminative power and invariance criteria.

In order to illustrate the impact of the detectors and matching approaches on the descriptor performance, different combinations of detectors and matching strategies are tested.

In the beginning, we have expected that better performances of our approach would be obtained within scenes depicting a large number of similar regions, textured scenes, and scenes reflecting complicated non-affine transformations.

Overall, the experimental results showed SSC-based descriptors to perform significantly above the expected performance. It clearly illustrated the effectiveness of the SSC information, under particular conditions, to turn impractical descriptors into well suited descriptors.

This has been demonstrated for geometric image transformations, rotation, scaling and viewpoint change. We also noticed the usefulness of our approach under imaging condition changes, is less but still outperforms other descriptors as we showed for SIFT-Based-SSC under image blur, illumination change, and JPEG compression.

Further, we observed how well the performance of the basic descriptor, cross-correlation is highly increased when including the SSC component as in cases of geometric transformations and image blur.

Besides, we retained that SIFT-Based-SSC performs best performance across all the comparisons, *i.e.*, for different detectors, matching strategies, and scene types.

According to different variants of tested SSC-based descriptor, the best gains for both the discriminative power and invariance are obtained mostly with SIFT-Based-SSC and CC-Based-SSC descriptors. However, SPIN-Based-SSC is better for image rotation and blur than for scale changes and JPEG compression.

We found the gains obtained with CC-Based-SSC under JPEG compression are much more than those of SIFT-Based-SSC. In addition, the best performance of SPIN-Based-SSC in terms of discriminative power is obtained for the textured scene of image rotation with descriptors computed for `hessian-laplace` where SIFT-Based-SSC and CC-Based-SSC are outperformed by SPIN-Based-SSC.

Regarding the detectors, the performance of SSC-based descriptors are mostly similar. However, descriptors computed for Hessian-based detectors perform some times better than those of Harris-based detectors. As example, for image rotations, scale changes and viewpoint changes.

This is because the Hessian-based detectors are more accurate than those of Harris-based, as noted by their authors [Mikolajczyk 05b, Mikolajczyk 05a].

According to matching strategies, the SSC-based approaches show to perform equally with both the `nearest-neighbor` and `threshold-based` algorithms whereas in some evaluations, the performance appears to be better with the `nearest-neighbor` than with the `threshold-based` matching method. This is mainly due to the higher precision of the `nearest-neighbor` algorithm since it is almost correct and selects only one match (the best below the threshold) and discarding others. In contrast with it, the `threshold-based` algorithm selects many matches and obviously many of them are false which leads to lower the discriminative power.

Furthermore, in many cases, SIFT-Based-SSC appeared to win largely the other descriptors while it gains the first rank far away from them. This is illustrated for image rotations, scale changes, and the challenging transformation of viewpoint changes in particular.

Tab. 7.1 reports the ranking of descriptors over all the conducted evaluations with respect to different image deformations. This ranking is established according to the discriminative power of large image deformations. The descriptors are computed on `hessian-affine` regions and then matched with the `nearest-neighbor` algorithm. Since it is the most significant ROC region, the ranking is mostly based on descriptor responses nearby the left-side region of ROC space. For some situations, the ranking is not accurate enough because some descriptors perform almost equally. For illumination change and JPEG compression, the ranking is given for descriptors computed on `hessian-laplace` support regions.

This table shows clearly how well the performance of SIFT descriptor is enhanced once the SSC information is added. In this spirit, we obtain the high-ranking of SIFT-Based-SSC even though SIFT performs less good like for image rotation in which it is ranked 5 and 6.

Besides, the table shows the performance of the simplest cross-correlation (CC) descriptor remarkably enhanced when the SSC component is incorporated. This can be observed, as example, for viewpoint changes under the textured scene containing a large number of similar regions (T*). Thus, while the CC is ranked 8, the CC-Based-SSC wins the 2nd rank, and therefore outperforms the competitive descriptors of GLOH and SC.

The most evaluations are performed for variants of SSC-based descriptors (*i.e.*, SIFT-Based-SSC, SPIN-Based-SSC, and CC-Based-SSC), computed with parameter values fixed roughly, we thus set the number of clusters to 25, the number of concentric shells to 12 and we adopted the `k-means` clustering strategy.

TAB. 7.1: Ranking of image feature descriptors based on discriminative power performances. Here, "S" and "T" denote the structured and textured scenes, respectively. In addition, "S*" and "T*" means the scene contains a large number of similar motifs. Moreover, "1^{+}" and " 1^{++}" reflect how much the gap in performances between the descriptors (*i.e.*, between #1 and #2 ranks) is large. Finally, "+" means the gap is quite large, while "++" indicates the gap is hugely large.

		SIFT	GLOH	SC	JLA	MOM	KOEN	CF	SPIN	PCA	CC	SIFT-Based-SSC	SPIN-Based-SSC	CC-Based-SSC
Image rotation	S	2	3	4	9	13	11	12	8	7	10	**1**	5	6
	S*	6	5	7	8	12	11	13	10	4	9	$\mathbf{1}^{+}$	**2**	**3**
	T	5	2	3	10	13	12	11	7	4	9	$\mathbf{1}^{+}$	6	8
Scale change	S	3	2	4	8	13	10	11	7	5	12	**1**	9	6
	T	4	3	2	9	13	10	12	8	5	10	$\mathbf{1}^{+}$	6	7
Viewpoint change	S	2	3	4	9	13	11	12	8	5	10	$\mathbf{1}^{+}$	7	6
	T	4	2	3	9	13	10	11	6	5	12	**1**	7	8
	T*	4	3	5	9	11	12	13	10	7	8	$\mathbf{1}^{++}$	6	**2**
Image blur	S	2	3	4	7	13	11	12	10	6	8	**1**	9	5
	T	2	4	3	8	13	10	12	9	5	11	**1**	6	7
Illumination change	S	2	3	4	8	9	11	10	12	5	6	**1**	13	7
JPEG compression	S	2	3	4	8	9	11	12	13	6	7	**1**	10	5

Even though its best performance in terms of the discriminative power and invariance, our approach is not without problems. It suffers from an excessive computation time, such as the best case of SIFT-Based-SSC where we reported an extra time of $\approx 50\%$ in addition to that needed to compute the local component SIFT.

To overcome this constraint, we advocate building the SSC component around a simple local signature designed specifically with the lower computational time and to obtain a SSC component which can compensate for lack of performance of the local signature.

This is mainly inspired by the cross-correlation which showed an enormous performance enhancement once the SSC information is added while it is much less computationally demanding than other competitive local signatures, like SIFT.

This problem seems interesting to be addressed since in addition the advantage to improve the computation time, it allows to figure out the best approach for constructing the local signature in such a way to take benefits of extra performance from SSC component contributions.

In other words, investigate the appropriate type of local signatures as well as their suited connections with the SSC component so as to get the best performance of the resulted descriptor. Precisely, we suggest to inspect the best approach to build the local and SSC components in such a manner to compensate for lack in performance from one to another. It also means examine other alternatives to generate the SSC information instead of those based on clustering method.

7.3 SSC-Based Approach in 3D-Domain

The usefulness of our SSC approach is also evaluated in 3D domain, in which a SSC-based descriptor is proposed to resolve a problem arising in matching and alignment of multiple range images.

Similar to 2D domain, our suggested 3D feature augmented descriptor is computed based on extracting semantic features using `k-means` clustering method. These are then exploited to generate a SSC component, SSC, which in turn, is combined with local information, LD, to obtain our full SSC-based feature descriptor, GD.

Thus, a novel descriptor robust to overlap and extremely discriminative is derived. The proposed SSC-based descriptor is effective yet simple to implement.

The performance of SSC-based feature descriptor is evaluated for matching and alignment of multiple range images by performing different evaluations.

We started by conducting pairwise matching and alignment test to compare the per-

formances of LD, SSC, and GD components. The purpose is figuring out how well the performance is enhanced when combining local and SSC information compared to those of LD and SSC used separately. To this aim, the LD is computed based on one information only.

Next, the performance of the SSC-based descriptor built around multiple information, *i.e.*, different geometric measures, is compared to those of the standard 3D descriptors of 3D Shape-Context and Spin-Image.

In the second evaluation, the performance of the proposed approach is used to perform an automatic multiple range image registration. Similar to above, the SSC-based descriptor, GD, is computed around two different source of information, which are the shape index and β-value geometric measures.

The reported evaluation results illustrate without doubt the best performance of our descriptor compared to those of 3D Shape-Context and Spin-Image. This holds on different evaluated models when the approximate pairwise registration (*i.e.*, RANSAC pre-alignment) obtained with our descriptor is still the best for the almost pre-alignments. Moreover, in many cases the 3D Shape-Context and Spin-Image failed completely to produce an accurate enough pre-alignment, which is required by any ICP-based alignment algorithm in order to converge.

The results also demonstrate the robustness of our descriptor against the overlap area, in sense that for low (*i.e.*, small and medium) overlaps, the pre-alignment errors reported with our descriptor are mostly smaller than those of other descriptors.

Moreover, the result of the automatic multiple range image registration experiment shows that the registered (*i.e.*, aligned) models obtained with our approach appear without any visible imperfections and are much more accurate compared to those obtained with 3D Shape-Context and Spin-Image.

We also introduced an approach to automatically estimate the overlap area. We used the scale for determining which range image overlap. Based on the best scale, we find the best pairs of range images to be registered to each other.

Finally, despite the best performance of the proposed approach obtained for both the 2D and 3D domains, its usefulness is still far from being completely investigated by the means of the present framework. We suppose no other similar approach for learning the local feature, seems existing in the current literature, as far as we know. For this reasons, and in addition to the above suggestion, it would be a good challenge to find other machine vision problems in which the approach might be more powerful and give better performance. In this context we think, for example, about object recognition and scene categorization.

Bibliography

[Abdel-Hakim 06] A.E. Abdel-Hakim & A.A. Farag. *CSIFT: A SIFT descriptor with color invariant characteristics*. Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 1978–1983. Ieee, 2006. 14, 26, 40

[Ashbrook 95] AP Ashbrook, NA Thacker, PI Rockett & CI Brown. *Robust recognition of scaled shapes using pairwise geometric histograms*. Proc. BMVC, pp. 503–512. Citeseer, 1995. 19

[Baumberg 00] A. Baumberg. *Reliable feature matching across widely separated views*. cvpr, pp. 1774. Published by the IEEE Computer Society, 2000. 10, 37

[Bay 06] H. Bay, T. Tuytelaars & L. Van Gool. *Surf: Speeded up robust features*. Computer Vision–ECCV 2006, pp. 404–417, 2006. 2, 17, 26, 38, 42

[Bay 08] H. Bay, A. Ess, T. Tuytelaars & L. Van Gool. *Speeded-up robust features (SURF)*. Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008. 17, 26, 61

[Belongie 02] S. Belongie, J. Malik & J. Puzicha. *Shape matching and object recognition using shape contexts*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 509–522, 2002. 2, 4, 8, 9, 19, 20, 26, 30, 40, 47, 63, 71

[Berg 01] A.C. Berg & J. Malik. *Geometric blur for template matching*. 2001. 3, 20, 22, 26

[Besl 92] P.J. Besl & H.D. McKay. *A method for registration of 3-D shapes*. IEEE Transactions on pattern analysis and machine intelligence, vol. 14, no. 2, pp. 239–256, 1992. 27, 157

[Blank 05] M. Blank, L. Gorelick, E. Shechtman, M. Irani & R. Basri. *Actions as Space-Time Shapes*. Proceedings of the Tenth IEEE International Conference on Computer Vision-Volume 2, pp. 1395–1402. IEEE Computer Society, 2005. 15, 32, 39

[Borgefors 84] G. Borgefors. *Distance transformations in arbitrary dimensions*. Computer Vision, Graphics, and Image Processing, vol. 27, no. 3, pp. 321–345, 1984. 15

[Bosch 06] A. Bosch, A. Zisserman & X. Munoz. *Scene classification via pLSA*. Computer Vision–ECCV 2006, pp. 517–530, 2006. 39, 40

[Bradley 97] P.S. Bradley, O.L. Mangasarian & W.N. Street. *Clustering via concave minimization*. Advances in neural information processing systems, pp. 368–374, 1997. 50

[Briechle 01] K. Briechle & U.D. Hanebeck. *Template matching using fast normalized cross correlation*. Proceedings of SPIE, vol. 4387, pp. 95. Spie, 2001. 13, 26

[Brown 03] M. Brown & D.G. Lowe. *Recognising panoramas*. Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 1218. Citeseer, 2003. 2

[Burghouts 09] G.J. Burghouts & J.M. Geusebroek. *Performance evaluation of local colour invariants*. Computer Vision and Image Understanding, vol. 113, no. 1, pp. 48–62, 2009. 40, 61

[Canny 86] J. Canny. *A computational approach to edge detection*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 6, pp. 679–698, 1986. 9, 19

[Carneiro 02] G. Carneiro & A. Jepson. *Phase-based local features*. ECCV 2002, pp. 282–296, 2002. 37

[Carneiro 03] G. Carneiro & A.D. Jepson. *Multi-scale phase-based local features*. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 1, pp. I–736. IEEE, 2003. 23

[Carneiro 04] G. Carneiro & A.D. Jepson. *Pruning local feature correspondences using shape context*. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3, pp. 16–19. IEEE, 2004. 4, 23

[Castellani 08] U. Castellani, M. Cristani, S. Fantoni & V. Murino. *Sparse points matching by combining 3D mesh saliency with statistical descriptors*. Computer Graphics Forum, vol. 27, pp. 643–652. Blackwell Publishing, 2008. 43

[Chen 03] D.Y. Chen, X.P. Tian, Y.T. Shen & M. Ouhyoung. *On visual similarity based 3D model retrieval*. Computer graphics forum, vol. 22, pp. 223–232. Wiley Online Library, 2003. 29, 30, 36

[Cheng 98] S.F. Cheng, W. Chen & H. Sundaram. *Semantic visual templates: linking visual features to semantics*. Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, pp. 531–535. IEEE, 1998. 50

[Chua 97] C.S. Chua & R. Jarvis. *Point signatures: A new representation for 3d object recognition*. International Journal of Computer Vision, vol. 25, no. 1, pp. 63–85, 1997. 33

[Csurka 04] G. Csurka, C. Dance, L. Fan, J. Willamowski & C. Bray. *Visual categorization with bags of keypoints*. Workshop on Statistical Learning in Computer Vision, ECCV, vol. 2004, 2004. 4, 44, 46

[Cui 09] Y. Cui, N. Hasler, T. Thormählen & H.P. Seidel. *Scale invariant feature transform with irregular orientation histogram binning*. Image Analysis and Recognition, pp. 258–267, 2009. 17, 26

[Cula 01] O.G. Cula & K.J. Dana. *Compact representation of bidirectional texture functions*. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pp. I–1041. IEEE, 2001. 37

[Curless 99] B. Curless. *From range scans to 3D models*. ACM SIGGRAPH Computer Graphics, vol. 33, no. 4, pp. 38–41, 1999. 28

[Dalal 06] N. Dalal, B. Triggs & C. Schmid. *Human detection using oriented histograms of flow and appearance*. Computer Vision–ECCV 2006, pp. 428–441, 2006. 15

[Daras 09] P. Daras & A. Axenopoulos. *A compact multi-view descriptor for 3D object retrieval*. Content-Based Multimedia Indexing, 2009. CBMI'09. Seventh International Workshop on, pp. 115–119. IEEE, 2009. 32, 36

[Deans 83] S.R. Deans. The radon transform and some of its applications. Wiley New York:, 1983. 15

[Duda 98] RO Duda, PE Hart & DG Stork. *Pattern Classification and Scene Analysis: Part I Pattern Classification*, 1998. 11

[Duda 01] R.O. Duda, P.E. Hart & D.G. Stork. Pattern classification. Wiley New York, 2001. 46, 156

[Everingham 07] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn & A. Zisserman. *The PASCAL visual object classes challenge 2007 (VOC2007) results*, 2007. 39

[Fawcett 04] T. Fawcett. *ROC graphs: Notes and practical considerations for researchers*. Machine Learning, vol. 31, no. HPL-2003-4, pp. 1–38, 2004. 61

[Fei-Fei 06] L. Fei-Fei, R. Fergus & P. Perona. *One-shot learning of object categories*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 594–611, 2006. 3, 40

[Fergus 03] R. Fergus, P. Perona & A. Zisserman. *Object class recognition by unsupervised scale-invariant learning*. 2003. 2

[Ferrari 04] V. Ferrari, T. Tuytelaars & L.V. Gool. *Simultaneous object recognition and segmentation by image exploration*. Computer Vision-ECCV 2004, pp. 40–54, 2004. 2

[Fischler 81] M. Fischler. *Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography*. Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981. 156

[Florack 91] L. Florack, BM ter Haar Romeny, JJ Koenderink & M. Viergever. *General intensity transformations and second order invariants*. Proc. Seventh Scandinavian Conf. Image Analysis, pp. 338–345, 1991. 40

[Florack 94] LMJ Florack, BM ter Haar Romeny, JJ Koenderink & MA Viergever. *General intensity transformations and differential invariants*. Journal of Mathematical Imaging and Vision, vol. 4, no. 2, pp. 171–187, 1994. 40

[Flusser 09] J. Flusser, T. Suk, B. Zitov & Inc Ebrary. Moments and moment invariants in pattern recognition. Wiley Online Library, 2009. 11

[Frank 05] A. Frank. *On Kuhn's Hungarian method-a tribute from Hungary*. Naval Research Logistics, vol. 52, no. 1, pp. 2–5, 2005. 156

[Freeman 91] W.T. Freeman, E.H. Adelson, Massachusetts Institute of Technology. Media Laboratory. Vision & Modeling Group. *The design and use of steerable filters*. IEEE Transactions on Pattern analysis and machine intelligence, vol. 13, no. 9, pp. 891–906, 1991. 10, 12, 26, 40, 63

[Frome 04] A. Frome, D. Huber, R. Kolluri, T. Bulow & J. Malik. *Recognizing objects in range data using regional point descriptors*. Lecture Notes in Computer Science, pp. 224–237, 2004. 4, 30, 36, 47, 158

[Fujita 92] I. Fujita, K. Tanaka, M. Ito & K. Cheng. *Columns for visual features of objects in monkey inferotemporal cortex*. Nature, vol. 360, no. 6402, pp. 343–346, 1992. 50

[Gabor 46] D. Gabor. *Theory of communication. Part 1: The analysis of information*. Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of, vol. 93, no. 26, pp. 429–441, 1946. 10

[Gabriel 03] P.F. Gabriel, J.G. Verly, J.H. Piater & A. Genon. *The state of the art in multiple object tracking under occlusion in video sequences*. Advanced Concepts for Intelligent Vision Systems, pp. 166–173. Citeseer, 2003. 8

[Geusebroek 00] J.M. Geusebroek. *Color and geometrical structure in images*. Appl. Microsc, pp. 116–119, 2000. 14

[Geusebroek 01] J.M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders & H. Geerts. *Color invariance*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1338–1350, 2001. 14

[Geusebroek 05] J.M. Geusebroek, G.J. Burghouts & A.W.M. Smeulders. *The Amsterdam library of object images*. International Journal of Computer Vision, vol. 61, no. 1, pp. 103–112, 2005. 40

[Gonzalez] R.C. Gonzalez & R.E. Woods. *Digital image processing. 1992*. Reading, Mass.: Addison-Wesley, vol. 16, no. 716, pp. 8. 11

[Harris 88a] C. Harris & M. Stephens. *A combined corner and edge detector*. Alvey vision conference, vol. 15, pp. 50. Manchester, UK, 1988. 42

[Harris 88b] C. Harris & M. Stephens. *A combined corner and edge detector*. Alvey Vision Conference, 1988. 65

[Heider 12] P. Heider, A. Pierre-Pierre, R. Li, R. Mueller & C. Grimm. *Comparing local shape descriptors*. The Visual Computer, pp. 1–11, 2012. 3

[Huang 06] C.R. Huang, C.S. Chen & P.C. Chung. *Contrast context histogram-a discriminating local descriptor for image matching*. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 4, pp. 53–56. IEEE, 2006. 18, 26

[Huang 08] C.R. Huang, C.S. Chen & P.C. Chung. *Contrast context histogram–An efficient discriminating local descriptor for object recognition and image matching*. Pattern Recognition, vol. 41, no. 10, pp. 3071–3077, 2008. 18, 19, 26

[Huber 03] D.F. Huber & M. Hebert. *Fully automatic registration of multiple 3D data sets*. Image and Vision Computing, vol. 21, no. 7, pp. 637–650, 2003. 2, 27, 156

[Huttenlocher 87] D.P. Huttenlocher & S. Ullman. *Object Recognition Using Alignment*. Proceedings, vol. 14, pp. 102. Computer Society Press of the IEEE, 1987. 37

[Johnson 97a] A.E. Johnson. *Spin-images: a representation for 3-D surface matching*. 1997. 9, 28, 36, 43

[Johnson 97b] A.E. Johnson & M. Hebert. *Surface registration by matching oriented points*. 3dim, pp. 121. Published by the IEEE Computer Society, 1997. 9

[Johnson 98] A.E. Johnson & M. Hebert. *Surface matching for object recognition in complex three-dimensional scenes*. Image and Vision Computing, vol. 16, no. 9-10, pp. 635–651, 1998. 28

[Johnson 99] A.E. Johnson & M. Hebert. *Using spin images for efficient object recognition in cluttered 3D scenes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 5, pp. 433–449, 1999. 4, 9, 26, 42, 44, 158

[Kadir 04] T. Kadir, A. Zisserman & M. Brady. *An affine invariant salient region detector*. Lecture Notes in Computer Science, pp. 228–241, 2004. 42

[Kazhdan 03] M. Kazhdan, T. Funkhouser & S. Rusinkiewicz. *Rotation invariant spherical harmonic representation of 3D shape descriptors*. Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pp. 156–164. Eurographics Association, 2003. 29, 31, 36

[Ke 04] Y. Ke & R. Sukthankar. *PCA-SIFT: A more distinctive representation for local image descriptors*. 2004. 2, 8, 14, 19, 26, 63

[Kläser 08] A. Kläser, M. Marszałek & C. Schmid. *A spatio-temporal descriptor based on 3D-gradients*. British Machine Vision Conference, pp. 995–1004. Citeseer, 2008. 15, 16, 26, 39

[Knopp 10] J. Knopp, M. Prasad, G. Willems, R. Timofte & L. Van Gool. *Hough transform and 3d surf for robust three dimensional classification*. Computer Vision–ECCV 2010, pp. 589–602, 2010. 43

[Koenderink 87] J.J. Koenderink & AJ Van Doorn. *Representation of local geometry in the visual system*. Biological cybernetics, vol. 55, no. 6, pp. 367–375, 1987. 10, 11, 26, 37, 40, 63

[Kokkinos 12] I. Kokkinos, M.M. Bronstein, R. Litman & A.M. Bronstein. *Intrinsic shape context descriptors for deformable shapes*. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 159–166. IEEE, 2012. 33

[Kuhn 05] HW Kuhn. *The Hungarian method for the assignment problem*. Naval Research Logistics, vol. 52, no. 1, 2005. 37

[Laptev 08] I. Laptev, M. Marszalek, C. Schmid & B. Rozenfeld. *Learning realistic human actions from movies*. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE, 2008. 15, 39

[Lazebnik 05] S. Lazebnik, C. Schmid & J. Ponce. *A sparse texture representation using local affine regions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1265–1278, 2005. 2, 9, 37, 40, 43, 63

[Lewis 95] JP Lewis. *Fast normalized cross-correlation*. Vision Interface, vol. 10, pp. 120–123. Citeseer, 1995. 13, 16, 26, 43, 63

[Liu 96] F. Liu & R.W. Picard. *Periodicity, directionality, and randomness: Wold features for image modeling and retrieval*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 18, no. 7, pp. 722–733, 1996. 37

[Lowe 99] D.G. Lowe. *Object recognition from local scale-invariant features*. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, pp. 1150–1157. Ieee, 1999. 8, 42

[Lowe 04] D.G. Lowe. *Distinctive image features from scale-invariant keypoints*. International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004. 2, 4, 13, 14, 15, 26, 43, 63

[MacQueen 67] J. MacQueen*et al.* *Some methods for classification and analysis of multivariate observations*. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 14. California, USA, 1967. 50

[Maji 09] S. Maji. *A Comparison of Feature Descriptors*. 2009. 3, 40, 61

[Makadia 06] A. Makadia, A.I.V. Patterson & K. Daniilidis. *Fully Automatic Registration of 3D Point Clouds*. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 1, pp. 1297–1304. IEEE Computer Society Washington, DC, USA, 2006. 2, 27

[Malassiotis 07] S. Malassiotis & M.G. Strintzis. *Snapshots: A Novel Local Surface Descriptor and Matching Algorithm for Robust 3D Surface Alignment*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 7, pp. 1285, 2007. 174

[Matas 04] J. Matas, O. Chum, M. Urban & T. Pajdla. *Robust wide-baseline stereo from maximally stable extremal regions*. Image and Vision Computing, vol. 22, no. 10, pp. 761–767, 2004. 42

[Matthews 02] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox & R. Harvey. *Extraction of visual features for lipreading*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 2, pp. 198–213, 2002. 50

[Mian 05] A.S. Mian, M. Bennamoun & R.A. Owens. *Automatic correspondence for 3d modeling: An extensive review*. International Journal of Shape Modeling, vol. 11, no. 2, pp. 253, 2005. 4

[Mian 06] A.S. Mian, M. Bennamoun & R.A. Owens. *A novel representation and feature matching algorithm for automatic pairwise registration of range images*. International Journal of Computer Vision, vol. 66, no. 1, pp. 19–40, 2006. 2, 30, 36

[Mian 10] A. Mian, M. Bennamoun & R. Owens. *On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes*. International Journal of Computer Vision, vol. 89, no. 2, pp. 348–361, 2010. 42

[Mikolajczyk 01] K. Mikolajczyk & C. Schmid. *Indexing based on scale invariant interest points*. 2001. 2, 38, 39, 65

[Mikolajczyk 04] K. Mikolajczyk & C. Schmid. *Scale & affine invariant interest point detectors*. International journal of computer vision, vol. 60, no. 1, pp. 63–86, 2004. 38, 40, 42, 52, 65

[Mikolajczyk 05a] K. Mikolajczyk & C. Schmid. *A performance evaluation of local descriptors*. IEEE transactions on pattern analysis and machine intelligence,

pp. 1615–1630, 2005. 2, 11, 13, 14, 26, 38, 40, 58, 59, 61, 62, 63, 112, 182

[Mikolajczyk 05b] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir & L.V. Gool. *A comparison of affine region detectors*. International Journal of Computer Vision, vol. 65, no. 1, pp. 43–72, 2005. 62, 65, 182

[Mikolajczyk 05c] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir & L.V. Gool. *A comparison of affine region detectors*. International journal of computer vision, vol. 65, no. 1, pp. 43–72, 2005. 38, 42, 43

[Moreels 07] P. Moreels & P. Perona. *Evaluation of features detectors and descriptors based on 3D objects*. International Journal of Computer Vision, vol. 73, no. 3, pp. 263–284, 2007. 38

[Mortensen 05] E.N. Mortensen, H. Deng & L. Shapiro. *A SIFT descriptor with global context*. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 184–190. IEEE, 2005. 4, 23, 25, 26

[Noldus 01] L.P.J.J. Noldus, A.J. Spink & R.A.J. Tegelenbosch. *EthoVision: a versatile video tracking system for automation of behavioral experiments*. Behavior Research Methods, vol. 33, no. 3, pp. 398–414, 2001. 8

[Novatnack 08] J. Novatnack & K. Nishino. *Scale-Dependent/Invariant Local 3D Shape Descriptors for Fully Automatic Registration of Multiple Sets of Range Images*. Proceedings of the 10th European Conference on Computer Vision: Part III, pp. 440–453. Springer-Verlag Berlin, Heidelberg, 2008. 32, 36, 157

[Ojala 02] Timo Ojala, Matti Pietikainen & Topi Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 7, pp. 971–987, 2002. 10

[Opelt 04] A. Opelt, M. Fussenegger, A. Pinz & P. Auer. *Weak hypotheses and boosting for generic object detection and recognition*. Computer Vision-ECCV 2004, pp. 71–84, 2004. 2

[Petitjean 02] S. Petitjean. *A survey of methods for recovering quadrics in triangle meshes*. ACM Computing Surveys, vol. 34, no. 2, 2002. 4, 44

[Picard 93] R.W. Picard, T. Kabir & F. Liu. *Real-time recognition with the entire Brodatz texture database*. Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on, pp. 638–639. IEEE, 1993. 37

[Poston 96] T. Poston & I. Stewart. Catastrophe theory and its applications, vol. 2. Dover Pubns, 1996. 10

[Prince 02] S.J.D. Prince, K. Xu & A.D. Cheok. *Augmented reality camera tracking with homographies*. Computer Graphics and Applications, IEEE, vol. 22, no. 6, pp. 39–45, 2002. 59

[Randen 99] T. Randen & J.H. Husoy. *Filtering for texture classification: A comparative study*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 21, no. 4, pp. 291–310, 1999. 37

[Renninger 04] L.W. Renninger & J. Malik. *When is scene identification just texture recognition?* Vision Research, vol. 44, no. 19, pp. 2301–2311, 2004. 37

[Richardson 02] M.K. Richardson & G. Keuck. *Haeckel's ABC of evolution and development*. Biological Reviews, vol. 77, no. 4, pp. 495–528, 2002. 55

[Schaffalitzky 02] F. Schaffalitzky & A. Zisserman. *Multi-view matching for unordered image sets*. Computer Vision, ECCV 2002, pp. 414–431, 2002. 2, 10, 26, 37, 63

[Schmid 97] C. Schmid & R. Mohr. *Local grayvalue invariants for image retrieval*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no. 5, pp. 530–535, 1997. 2, 37

[Schmid 01] C. Schmid. *Constructing models for content-based image retrieval*. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 2, pp. II–39. IEEE, 2001. 37

[Schuldt 04] C. Schuldt, I. Laptev & B. Caputo. *Recognizing Human Actions: A Local SVM Approach*. Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3-Volume 03, pp. 32–36. IEEE Computer Society, 2004. 15, 39

[Scovanner 07] P. Scovanner, S. Ali & M. Shah. *A 3-dimensional sift descriptor and its application to action recognition*. Proceedings of the 15th international conference on Multimedia, pp. 357–360. ACM, 2007. 32, 36

[Se 02] S. Se, D. Lowe & J. Little. *Global localization using distinctive visual features*. Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on, vol. 1, pp. 226–231. IEEE, 2002. 2

[Shilane 04] P. Shilane, P. Min, M. Kazhdan & T. Funkhouser. *The princeton shape benchmark*. Shape Modeling Applications, 2004. Proceedings, pp. 167–178. Ieee, 2004. 30

[Sivic 03] J. Sivic & A. Zisserman. *Video Google: A text retrieval approach to object matching in videos*. 2003. 2

[Smeaton 06] A.F. Smeaton, P. Over & W. Kraaij. *Evaluation campaigns and TRECVid*. Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 321–330. ACM, 2006. 40

[Snoek 06] C.G.M. Snoek, M. Worring, J.C. Van Gemert, J.M. Geusebroek & A.W.M. Smeulders. *The challenge problem for automated detection of 101 semantic concepts in multimedia*. Proceedings of the 14th annual ACM international conference on Multimedia, pp. 421–430. ACM, 2006. 40

[Strecha 08] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua & U. Thoennessen. *On benchmarking camera calibration and multi-view stereo for high resolution imagery*. 2008. 39

[Sun 09] J. Sun, M. Ovsjanikov & L. Guibas. *A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion*. Computer Graphics Forum, vol. 28, pp. 1383–1392. Wiley Online Library, 2009. 43

[Tabbone 06] S. Tabbone, L. Wendling & J.P. Salmon. *A new shape descriptor defined on the Radon transform*. Computer Vision and Image Understanding, vol. 102, no. 1, pp. 42–51, 2006. 15, 26

[Tang 12] S. Tang & A. Godil. *An evaluation of local shape descriptors for 3D shape retrieval*. Arxiv preprint arXiv:1202.2368, 2012. 3, 33

[Theodoridis 06] S. Theodoridis & K. Koutroumbas. *Clustering: basic concepts*. Pattern Recognition,, pp. 483–516, 2006. 50

[Thrun 00] S. Thrun, W. Burgard & D. Fox. *A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping*. Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on, vol. 1, pp. 321–328. IEEE, 2000. 8

[Thrun 02] S. Thrun. Robotic mapping: A survey. Morgan Kaufmann, 2002. 8

[Toews 09] M. Toews & W. Wells. *SIFT-Rank: Ordinal description for invariant feature correspondence*. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 172–177. IEEE, 2009. 16, 26

[Tola 08] E. Tola, V. Lepetit & P. Fua. *A fast local descriptor for dense matching*. Proc. CVPR, vol. 2. Citeseer, 2008. 15, 26, 38, 39

[Tola 09] E. Tola, V. Lepetit & P. Fua. *Daisy: An efficient dense descriptor applied to wide-baseline stereo*. IEEE transactions on pattern analysis and machine intelligence, pp. 815–830, 2009. 39

[Tombari 10] F. Tombari, S. Salti & L. Di Stefano. *Unique signatures of histograms for local surface description*. Computer Vision–ECCV 2010, pp. 356–369, 2010. 32

[Tombari 11] F. Tombari, S. Salti & L. Di Stefano. *A combined texture-shape descriptor for enhanced 3D feature matching*. Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 809–812. IEEE, 2011. 32, 33, 36

[Tombari 12] F. Tombari, S. Salti & L. Di Stefano. *Performance Evaluation of 3D Keypoint Detectors*. International Journal of Computer Vision, pp. 1–23, 2012. 42

[Torralba 07] A. Torralba, K.P. Murphy & W.T. Freeman. *Sharing visual features for multiclass and multiview object detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 854–869, 2007. 50

[Trucco 06] E. Trucco & K. Plakas. *Video tracking: a concise survey*. Oceanic Engineering, IEEE Journal of, vol. 31, no. 2, pp. 520–529, 2006. 8

[Tuytelaars 99] T. Tuytelaars, L. Van Gool, L. D'haene & R. Koch. *Matching of affinely invariant regions for visual servoing*. Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on, vol. 2, pp. 1601–1606. IEEE, 1999. 42

[Tuytelaars 04] T. Tuytelaars & L. Van Gool. *Matching widely separated views based on affine invariant regions*. International journal of computer vision, vol. 59, no. 1, pp. 61–85, 2004. 2, 42

[Tuzel 06] O. Tuzel, F. Porikli & P. Meer. *Region covariance: A fast descriptor for detection and classification*. Computer Vision–ECCV 2006, pp. 589–600, 2006. 15, 26

[Ullman 02] S. Ullman, M. Vidal-Naquet & E. Sali. *Visual features of intermediate complexity and their use in classification*. nature neuroscience, vol. 5, no. 7, pp. 682–687, 2002. 50

[Unnikrishnan 08] R. Unnikrishnan & M. Hebert. *Multi-scale interest regions from unorganized point clouds*. Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp. 1–8. IEEE, 2008. 42

[Van De Sande 09] K.E.A. Van De Sande, T. Gevers & C.G.M. Snoek. *Evaluating color descriptors for object and scene recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1582–1596, 2009. 39

[Van De Weijer 06] J. Van De Weijer, T. Gevers & A.D. Bagdanov. *Boosting color saliency in image feature detection*. IEEE transactions on pattern analysis and machine intelligence, pp. 150–156, 2006. 39

[Van Gool 96] L. Van Gool, T. Moons & D. Ungureanu. *Affine/photometric invariants for planar intensity patterns*. Computer Vision, ECCV'96, pp. 642–651, 1996. 11, 26, 40, 63

[Van Nguyen 11] H. Van Nguyen & F. Porikli. *Concentric ring signature descriptor for 3D objects*. Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 2893–2896. IEEE, 2011. 32, 34, 36

[Varma 02] M. Varma & A. Zisserman. *Classifying images of materials: Achieving viewpoint and illumination independence*. Computer Vision, ECCV 2002, pp. 255–271, 2002. 37

[Varma 07] M. Varma & D. Ray. *Learning the Discriminative Power-Invariance Trade-Off*. 2007. 52

[Vranic 01a] D.V. Vranic & D. Saupe. *3D shape descriptor based on 3D Fourier transform*. Proceedings of the EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services (ECMCS 2001), Budapest, Hungary, 2001. 28, 30, 36

[Vranic 01b] D.V. Vranic, D. Saupe & J. Richter. *Tools for 3D-object retrieval: Karhunen-Loeve transform andspherical harmonics*. 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pp. 293–298, 2001. 27

[Vranic 05] D.V. Vranic. *Desire: a composite 3d-shape descriptor*. Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp. 4–pp. IEEE, 2005. 30, 36

[Vyas] N. Vyas. *Using Geometric Blur for Point Correspondence*. 26

[Winder 07] S.A.J. Winder & M. Brown. *Learning local image descriptors*. 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, 2007. 15

[Wong 07] S.F. Wong & R. Cipolla. *Extracting Spatiotemporal Interest Points using Global Information*. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pp. 1–8. IEEE, 2007. 15, 39

[Wu 04] B.Y. Wu & K.M. Chao. Spanning trees and optimization problems, vol. 19. CRC Press, 2004. 164

[Yoo 09] J.C. Yoo & T.H. Han. *Fast normalized cross-correlation*. Circuits, Systems, and Signal Processing, vol. 28, no. 6, pp. 819–843, 2009. 13, 26

[Zabih 94] R. Zabih & J. Woodfill. *Non-parametric local transforms for computing visual correspondence*. Computer Vision, ECCV'94, pp. 151–158, 1994. 10, 26

[Zaharescu 09] A. Zaharescu, E. Boyer, K. Varanasi & R. Horaud. *Surface feature detection and description with applications to mesh matching*. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 373–380. Ieee, 2009. 32, 42

[Zaharia 01] T. Zaharia & F. Prêteux. *3D shape-based retrieval within the MPEG-7 framework*. Proc. SPIE Conf. on Nonlinear Image Processing and Pattern Analysis XII, vol. 4304, pp. 133–145, 2001. 29, 36

[Zahn 72] C.T. Zahn & R.Z. Roskies. *Fourier descriptors for plane closed curves*. Computers, IEEE Transactions on, vol. 100, no. 3, pp. 269–281, 1972. 29

[Zhang 07] J. Zhang, M. Marszalek, S. Lazebnik & C. Schmid. *Local features and kernels for classification of texture and object categories: A comprehensive study*. Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, pp. 13–13. IEEE, 2007. 39

[Zhong 09] Y. Zhong. *Intrinsic shape signatures: A shape descriptor for 3d object recognition*. Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 689–696. IEEE, 2009. 42