Available online at www.sciencedirect.com
ScienceDirect
Graphical Models xxx (2006) xxx–xxx
Graphical Models
www.elsevier.com/locate/gmod

SVD-matching using SIFT features

Elisabetta Delponte *, Francesco Isgrò, Francesca Odone, Alessandro Verri

DISI, Università di Genova, Via Dodecaneso 35, Genova I-16146, Italy

Received 31 January 2006; received in revised form 31 March 2006; accepted 7 July 2006

Abstract

The paper tackles the problem of feature points matching between pair of images of the same scene. This is a key problem in computer vision. The method we discuss here is a version of the SVD-matching proposed by Scott and Longuet-Higgins and later modified by Pilu, that we elaborate in order to cope with large scale variations. To this end we add to the feature detection phase a keypoint descriptor that is robust to large scale and view-point changes. Furthermore, we include this descriptor in the equations of the proximity matrix that is central to the SVD-matching. At the same time we remove from the proximity matrix all the information about the point locations in the image, that is the source of mismatches when the amount of scene variation increases. The main contribution of this work is in showing that this compact and easy algorithm can be used for severe scene variations. We present experimental evidence of the improved performance with respect to the previous versions of the algorithm.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Point matching; Spectral methods; Scale invariant features

1. Introduction

Finding correspondences between feature points is one of the keystones of computer vision, with application to a variety of problems. For this reason it has been tackled since the old days of computer vision research [44,31]. Automatic feature matching is often an initialisation procedure for more complex tasks, such as fundamental matrix estimation, image mosaicing, object recognition, and three-dimensional point clouds registration.

In this paper we consider the case when the epipolar geometry is not known, and then the corresponding point can be anywhere in the image. Also, we are interested in dealing with the correspondence problem as the baseline grows.

Classical approaches to point matching with unknown geometry assume a short baseline, and they are usually based on correlation (see, for instance, [8]). It is well known that correlation-based approaches suffer from view-point changes and do not take into account the global structure of the image. On this respect an elegant approach, falling in the family of spectral based methods, is due to Scott and Longuet-Higgins [36].

Spectral graph theory [5] aims to characterise the global structural properties of graphs using

* Corresponding author. Fax: +39 010 353 6699.
E-mail addresses: delponte@disi.unige.it (E. Delponte), isgro@na.infn.it (F. Isgrò), odone@disi.unige.it (F. Odone), verri@disi.unige.it (A. Verri).

1524-0703/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.gmod.2006.07.002

the eigenvalues and eigenvectors of an affinity matrix. Recently, it has been applied to a variety of computer vision and pattern matching problems, including point and shapes matching, and image segmentation [43,37,4,35,45]. The point matching method by Scott and Longuet-Higgins is based on computing a proximity matrix that depends on the distance between points belonging to the two images. The method performs well on synthetic images, but it is sensitive to the noise that affects points detection and localisation in real images. More recently, Pilu [29] suggested a modification of the method through a correspondence matrix encoding both proximity and similarity information. The similarity is computed as the normalised correlation between the points neighbourhoods. Experimental evidence shows that the method performs very well on stereo pairs, but performance drops as the baseline grows.

We claim that the reason for this behaviour is due to the feature descriptor adopted, more than to a limit of the algorithm. In this paper we propose a variant of the SVD-matching that uses scale invariant keypoints to tackle both scale changes and view-point changes.

Scale invariant features (often referred to as SIFT) were first proposed in [23] and attracted the attention of the computer vision community for their tolerance to scale, rotation, and view-point variations. A comparative study of many local image descriptors [26] shows the superiority of SIFT with respect to other feature descriptors for the case of several local transformations.

In our method we first locate keypoints with an affine invariant Harris corner detector [27], and we compute a SIFT description. We then build a correspondence matrix which is based on the distance between SIFT descriptors, discarding proximity information entirely. We present an extensive experimental analysis, judging the performances of our approach with respect to the original SVD-matching [29], a matching based on the Euclidean distance between SIFT [23], and to a previous version of this work that was using both SIFT similarity and proximity information [7].

The experimental results show that including SIFT point descriptors in the SVD-matching improves the performance with respect to the past versions of this algorithm. In particular it returns good results for scale changes, and large view-point variations. The current version still does not cope with wide-baselines. These conclusions are supported by an extensive experimental evaluation on different typologies of image data.

The paper is organised as follows. Section 2 gives and overview of the state of the art on image matching, also with a reference to other spectral-based methods. In Section 3 we recall the SVD-matching algorithm, while in Section 4 we describe the modified SVD-matching. Section 5 is left to the experimental analysis and to the comparative evaluation. A final discussion, in Section 6, ends the paper.

2. Related work

The state of the art on algorithms for image matching is vast. It is common practise to distinguish between feature-based methods and direct methods. The former rely on first acquiring image meaningful features and then matching them, producing sparse matches. Direct methods try to find matches over all image positions. The results are dense disparity maps, less reliable in flat areas. Direct methods usually assume a small view-point change, and they are often applied to stereo and motion estimation. In this paper we focus on feature-based methods: this section reviews the main contributions to this topic, mainly on dealing with viewpoint and scale changes. The section ends with a brief overview of spectral methods applied to matching.

2.1. Local interest points

Early works on matching images with salient features were based on using small amounts of local information to describe meaningful *keypoints*, such as corners [28,17]. Harris [16] showed that corners were efficient for tracking and estimating structure from motion. Applications to these fields were extended later by Shi and Tomasi [38]. In early works corners were simply represented using correlation windows centred around the keypoint. When matching these keypoints the underlying assumption is that no relevant changes occurred in illumination and scale.

Instead, every object in an image assumes a different meaning if observed at different scales, or under different illumination conditions. Another source of changes in appearance is view-point variation. These issues have been extensively studied in the last decades. For what concerns scale invariance many contributions appeared in the past [6,20,39], in particular it is worth mentioning the *scale-space* approach [20]. Scale-space is an effective framework to handle objects at different scales: an image is represented at different resolution levels, and the description obtained is not a simple random suppression of details, but it is a well-defined process that guarantees linearity and scale-space invariance. Detecting local features in a scale-space representation allows us to estimate the keypoint scale as well as its position. A great part of local keypoints that have been recently proposed follow this approach. A foremost aspect of the scale-space approach is that there are methods [19] that automatically choose the appropriate resolution level, discarding useless information. Thus, scale invariant features can be obtained by applying, for instance, a Harris detector at different scales, and then estimating the most meaningful scale for each keypoint. Alternatively, features can be localised directly on a scale-space structure, searching local maxima both on space and scale. The latter approach has been followed by Lowe in designing his Difference of Gaussians (DoG) feature detector [23].

Among the many recent works populating the literature on keypoint detection, it is worth mentioning the *scale and affine invariant interesting points* recently proposed by Mikolajczyk and Schmid [27], as they appear to be among the most promising keypoint detectors to date. The detection algorithm can be sketched as follows: first Harris corners are detected at multiple scales, then points at which a local measure of variation is maximal over scale are selected. This provides a set of distinctive points at the appropriate scale. Finally, an iterative algorithm modifies location, scale, and neighbourhood of each point and converges to affine invariant points.

In many application domains it has been shown that an efficient keypoint localisation should be associated to a feature description less sensitive to view-point changes than grey levels [46,11], and possibly embed invariance to rotation, scale, or illumination changes. One of the pioneering works in this direction is due to Schmid and Mohr [32]. They showed that local feature matching could be applied effectively to image recognition if a more robust feature description was used: they located the keypoints with a Harris detector, and then used a rotationally invariant descriptor of the local image region centred at the keypoint. From the same research group we get a comparative study on the effectiveness of the various invariant feature descriptors proposed so far [33]. Here it is shown that *SIFT (scale invariant feature transform)* [22] lead to excellent performances compared to other existing approaches. SIFT description is computed as follows: once a keypoint is located and its scale has been estimated, one or more orientations are assigned to it based on local image gradient direction around the keypoint. Then, image gradient magnitude and orientation are sampled around the keypoint, using the scale of the keypoint to select the level of Gaussian blur. The gradient orientations obtained are rotated with respect to the keypoint orientation previously computed. Finally, the area around the keypoint is divided in sub-regions, each of which is associated an orientations histogram weighted with the magnitude. This approach has been suggested to the author by a model of biological vision [9].

Other local keypoint descriptors can be found in the recent literature: Baumberg [2] propose a matching technique based on the Harris corner detector and a description based on the Fourier–Mellin transform to achieve invariance to rotation. Harris corners are also used in [1], where rotation invariance is obtained by a hierarchal sampling that starts from the direction of the gradient. Matas et al. [25] introduce the concept of maximally stable extremal region to be used for robust matching. These region's are connected components of pixels which are brighter or darker than pixels on the region's contour, they are invariant to affine and perspective transform, and to monotonic transformation of image intensities.

2.2. Matching with large or wide-baslines

It is well known that a major source of appearance variation is view-point change. This variation becomes more challenging to model as the distance between observation points (i.e., the baseline) grows. This section reviews some methods considering this issue.

Early applications to local image matching were stereo and short-range motion tracking. Zhang et al. showed that it was possible to match Harris corners over a large image range, with an outlier removal technique based on a robust computation of the fundamental matrix and the elimination of the feature pairs that did not agree with the solution obtained [47]. Later on, the invariant features described above were extensively studied as they guaranteed some degree of flexibility with respect to view-point change. Recently, many works on extending local features to be invariant to affine transformations have been proposed, including a variant of SIFT [3].

Tuytelaars and Van Gool [42] deal with wide-baseline matching extracting image region's around corners, where edges provide orientation and skew information. They also address scale variation by computing the extrema of a 2D affine invariant function; as a descriptor they use generalised colour moments. The actual matching is done using the Mahalanobis distance. In a more recent work [10] they establish wide-baseline correspondences among unordered multiple images, by first computing pairwise matches, and then integrating them into feature tracks each representing a local patch of the scene. They exploit the interplay between the tracks to extend matching to multiple views. A method based on automatic determination of local neighbourhood shapes is presented in [12], but it only works for image areas where stationary texture occurs.

An alternative approach for determining feature correspondences relies on prior knowledge on the observed scene, for instance in knowing the epipolar geometry of two or more views [34]. Georgis et al. [13] assume that projections of four corresponding non coplanar points at arbitrary positions are known. Pritchett and Zissermann [30] use local homographies determined by parallelogram structures or from motion pyramids. Lourakis et al [21] present a method based on the assumption that the viewed scene contains two planar surfaces and exploits the geometric constraints derived by this assumption. The spatial relation between the features in each images, together with appearance, is used in [40].

Recently a simple ordering constraint that can reduce the computational complexity for wide-basline matching, for the only case of approximately parallel epipolar lines, has been proposed in [24].

2.3. Spectral analysis for point matching

Spectral graph analysis aims at characterising the global properties of a graph using the eigenvalues and the eigenvectors of the graph adjacency matrix [5]. Recently this subject has found a number of applications to classical computer vision problems, including point matching, segmentation, line grouping, shape matching [43,37,4,35,45]. In this section we review some works on point matching with spectral analysis.

Most of these contributions are based on the so called *proximity* or *affinity* matrix, that is a continuous representation of the adjacency matrix: instead than being set to 0 or 1, the matrix elements are weights that reflect the strength of a pair relation (in terms of proximity or sometimes similarity). Usually the proximity matrix is defined as:

$$G_{ij} = \mathrm{e}^{-r_{ij}^2/2\sigma^2} \tag{1}$$

with σ a free parameter, and r_{ij} is a distance between points x_i and x_j computed with an appropriate affinity.

Scott and Longuet-Higgins [36] give one of the most interesting and elegant contributions to this topic, that we will describe in Section 3. One of the first applications of spectral analysis to point matching is due to Umeyama [43]. The author presents an SVD method for finding permutations between the adjacency matrixes of two graphs. If the graphs have the same size and structure of the edges the method is able to find correspondences between the nodes of the graph. Shapiro and Brady [37] propose a method that models the content of each image by means of an intra-image point proximity matrix, and then evaluates the similarity between images by comparing the matrixes. The proximity matrixes are built using a Gaussian weighting function, as in Eq. (1). For each proximity matrix, a modal matrix (a matrix the columns of which are eigenvectors of the original matrix) is built. Each row of the modal matrix represents one point of the corresponding image. The authors find the correspondences by comparing the rows of the two modal matrixes, using a binary decision function based on the Euclidean distance. Carcassoni and Hancock [4] propose a variant of this approach that changes the original method in three different ways. First, the evaluation of proximity matrixes are based on other weighting functions, including a sigmoidal and an Euclidean weighting function; second, the use of robust methods for comparing the modal matrixes; third, an embedding of the correspondence process within a graph matching EM algorithm. Experiments reported in the paper show that the latter contribution is useful to overcome structural errors, including the deletion or insertion of points. The authors also show that the Gaussian weighting function performs worst than the other weighting functions evaluated.

3. SVD-matching

In this section we summarise the algorithms proposed in [36] and [29] upon which we base our matching technique. Scott and Longuet-Higgins [36], getting some inspiration from structural chemistry, were among the first to use spectral methods for image matching. They show that, in spite of the well-known combinatorics complexity of finding feature correspondences, a reasonably good solu-

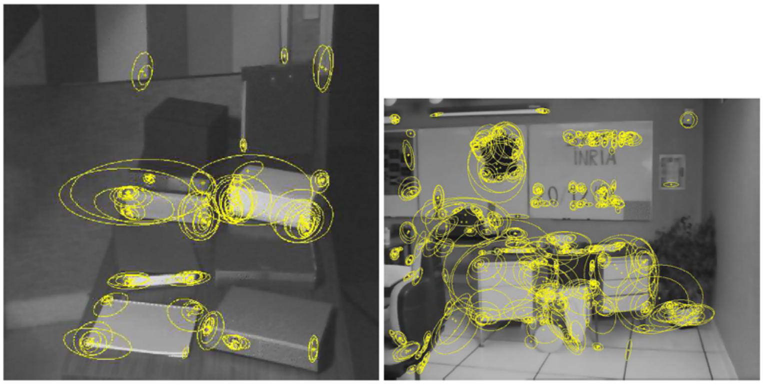

Fig. 1. Examples of features extracted. The ellipse around the feature points represents the support area of the feature.

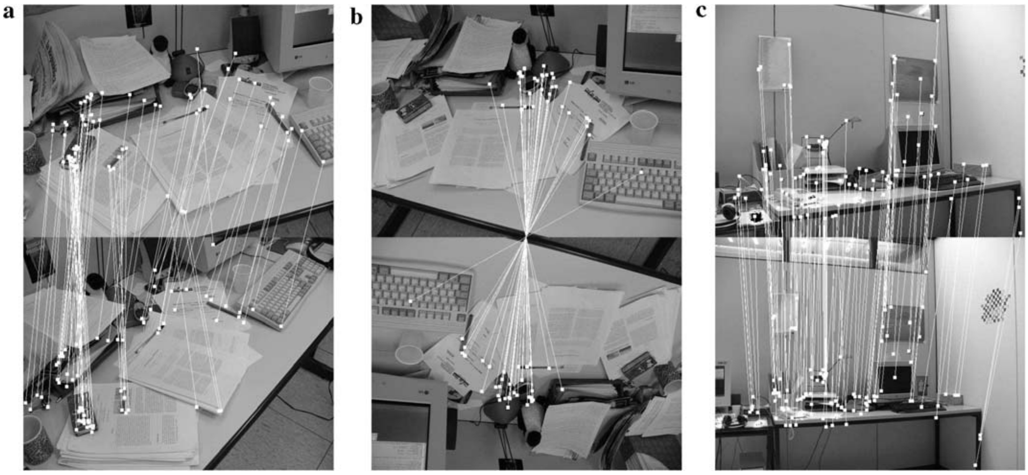

Fig. 2. Matches determined for stereo pairs of a desk. (a) A reasonable level of scene variation. We could notice only one wrong match between the wall and the corner of the screen. (b) The second image is a synthetic rotation of the first one. No wrong matches have been determined. (c) Scale variation, wrong matches on the edge of the table.

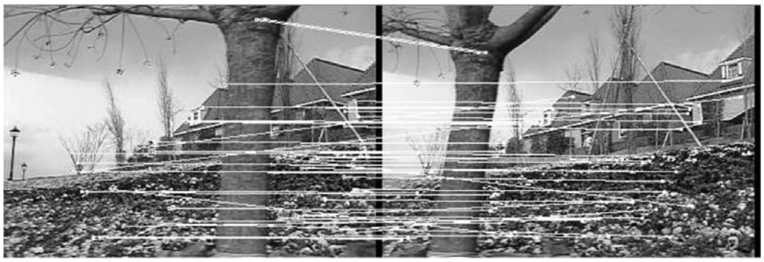

Fig. 3. Matches determined for a large baseline stereo pairs. Only 2–3 wrong matches are determined.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

tion can be achieved through the singular value decomposition of the proximity matrix of Eq. (1) followed by a simple manipulation of the eigenvalues. As pointed out in [29] their algorithm is rooted into the solution of the subspace rotation problem known as orthogonal Procrustes problem (see [14] for details).

Let A and B be two images, containing m and n features respectively (A_i, $i = 1,\ldots,m$, and $B_j, j = 1,\ldots,n$). The goal is to determine two subsets of the two sets of points that can be put in a one to one correspondence. In the original algorithm proposed by Longuet-Higgins, the main assumption was that the two images were taken from close points of view, so that the corresponding points had similar image coordinates.

The algorithm consists of three steps:

1. Build a proximity matrix **G**, where each element is computed according to Eq. (1). Let $r_{ij} = \|A_i - B_j\|$ be the Euclidean distance between the two points, when considering them in the same reference plane. The parameter σ controls the degree of interactions between features, where a small σ enforces local correspondences, while a bigger σ allows for more distant interactions. The elements of **G** range from 0 to 1, with higher values for closer points.

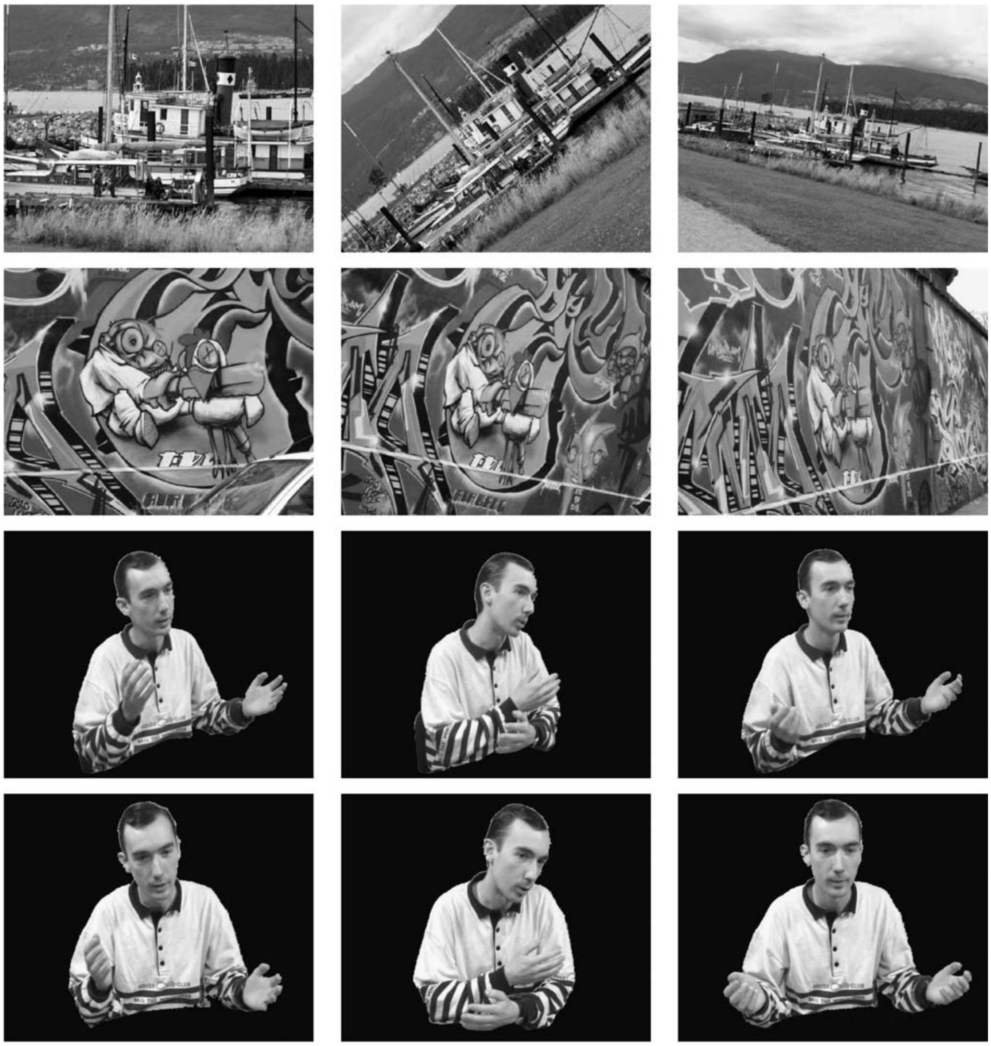

Fig. 4. Sample from the sequences used for the experiments presented in this paper. First row: 1st, 3rd and 5th frame of the *Boat* sequence. Second row: 1st, 3rd and 5th frame of the *Graf* sequence. Third and fourth rows: left and right views, respectively, of the 1st, 16th and 30th frame of the stereo sequence.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

2. Compute the Singular Value Decomposition for **G**: $\mathbf{G} = \mathbf{VDU}^\top$.
3. Compute a new correspondence matrix **P** by converting diagonal matrix **D** to a diagonal matrix **E** where each element D_{ii} is replaced with a 1: $\mathbf{P} = \mathbf{VEU}^\top$.

The algorithm is based on the two principles of proximity and exclusion, that is, corresponding points must be close, and each point can have one corresponding point at most. The idea is to obtain from the similarity matrix **G** a matrix **L** such that the entry *ij* is 1 if *i* and *j* are corresponding points, 0 otherwise. The matrix **P** computed by the algorithm is orthogonal (in the sense that the rows are mutually orthogonal), as all the singular values are 1, and it is the orthogonal matrix closest to the proximity **G**. Because of the orthogonality, if the parameter σ is chosen properly, **P** enhances good pairings, as its entries have properties close to those of the ideal matrix **L**. Following this idea the algorithms establishes a correspondence between the points *i* and *j* if the entry P_{ij} is the largest element in row *i* and the largest element in row *j*.

In the case of real images, point localisation is affected by noise and keypoint detection is unstable—keypoints may be detected or not depending on the viewing angle. The algorithm presented in [36] was working well on synthetic data, but performance started to fall down when moving to real images. Pilu [29] argues that this behaviour could be taken care of by evaluating local image similarities. He adapts the proximity matrix in order to take into account image intensity as well as geometric properties. The modified matrix appears as follows:

$$G_{ij} = \frac{C_{ij} + 1}{2} \mathrm{e}^{-r_{ij}^2/2\sigma^2} \tag{2}$$

where the term C_{ij} is the normalised correlation between image patches centred in the feature points.

In [29] experimental evidence is given that the proposed algorithm performs well on short baseline stereo pairs. In fact the performance falls when the baseline increases. It is our target to show that the reason for this behaviour is in the feature descriptor chosen and is not an intrinsic limit of the algorithm.

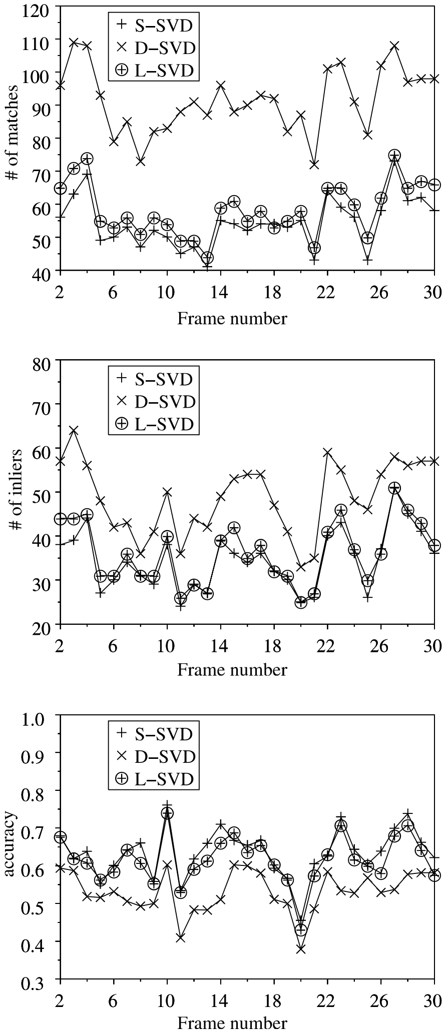

Fig. 6. Comparison with other weighting functions: results for the 30-frames stereo sequence. The baseline is fixed for all the stereo pairs, and the correspondences are computed for each stereo frame of the sequence. Top: total number of matches detected. Middle: number of correct matches. Bottom: accuracy of the method.

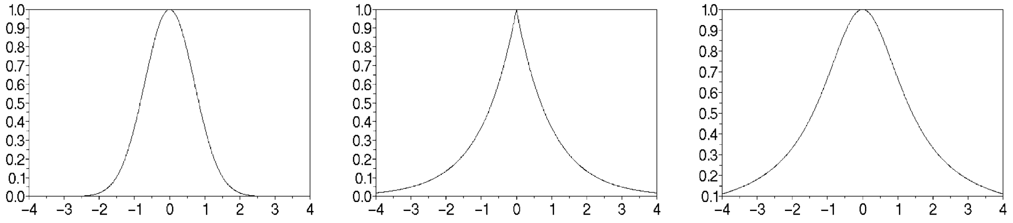

Fig. 5. The different weighting functions used. Left: Gaussian. Middle: Double-exponential. Right: Lorentzian.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

4. SVD-matching using SIFT

In this section we discuss the use of the SIFT descriptor in the SVD-matching algorithm. As mentioned in the previous section SVD-matching presented in [29] does not perform well when the baseline starts to increase. The reason for this behaviour is in the feature descriptor adopted. The original algorithm uses the grey level values in a neighbourhood of the keypoint. As pointed out in Section 2 this description is too sensitive to changes in the view-point, and more robust descriptor have been introduced so far.

A comparative study of the performance of various feature descriptors [26] showed that the SIFT descriptor is more robust than others with respect to rotation, scale changes, view-point change, and local affine transformations. The quality of the results decrease in the case of changes in the illumination. In the same work, cross-correlation between the image grey levels returned unstable performance, depending on the kind of transformation considered. The considerations above suggested the use of a SIFT descriptor, instead of grey levels. The descriptor is associated to scale and affine invariant interest points [27], briefly sketched in Section 2. Some examples of such keypoints are shown in Fig. 1.

In a previous version of this work [7] we left the matrix **G** in Eq. (2) unchanged in its form, but C_{ij} was the cross-correlation between SIFT descriptors. This straightforward modification improves the performance of the SVD-matching, and also gives better results, in terms of number of points correctly matched, with respect to the SIFT distance used for the experiments reported in [26]. However the matrix terms are still strongly dependent on the distance on the image plane between feature points, causing a large number of mismatches when the distance between points increases. For this reason we decided to switch back to the original form of the **G** matrix, with

$$G_{ij} = \mathrm{e}^{-r_{ij}^2/2\sigma^2} \tag{3}$$

where r_{ij} is now the distance between the feature descriptors in the SIFT space.

In order to reduce the number of mismatches even further we also added a constraint on the entry P_{ij} for determining the correspondence between points i and j. Let a_{ij_1} and a_{ij_2} being, respectively, the largest and second largest elements in row i, and b_{i_1j} and b_{i_2j} the largest and second largest elements in column j. We say that i and j are corresponding points if

1. $j_1=j$ and $i_1=i$

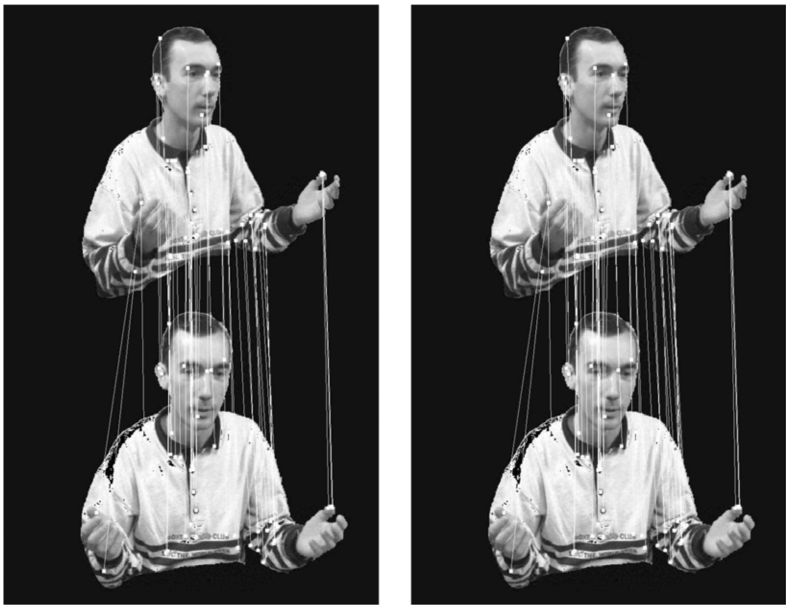

Fig. 7. Comparison of different weighting functions: results for the 30-frames stereo sequence. Correct matches between the left (top) and right (bottom) 27th frames. Left; S-SVD Right: L-SVD. The results for D-SVD are in Fig. 12.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

2. $0.6a_{ij_1} \geqslant a_{ij_2}$ and $0.6b_{ij_1} \geqslant b_{ij_2}$

In plain words it still needs to be the largest element in row *i* and column *j*, but also the largest by far.

5. Experimental results

In this section we report some experiments carried out on different image pairs and sequences. First we show some of the matches returned by our algorithm on few image pairs. Then we attempt a more quantitative analysis of the performance of our algorithm on short image sequences.

5.1. Experiments on image pairs

The first lot of experiments that we show refers to results on image pairs of two different scenes returned by the algorithm proposed in this paper.

In Figs. 2(a) and (b) we show all the matches determined on two pairs of images of a desk scene. The first one presents a reasonable level of scene variation, whereas the latter is a synthetic rotation of the first image. We spotted only a wrong match in Fig. 2(a). The last image pair is relative to a studio scene with scale variation. The result is shown in Fig. 2(c). Our visual inspection of the results determined only few wrong matches between points on the border of the table.

In Fig. 3 we show the matches determined on a large baseline stereo pair. A visual inspection could spot no more than three wrong matches.

5.2. Comparative experiments

We performed different comparative experiments. The first group of experiments focuses on proximity matrixes built in the descriptor space as for the one given in (3), that uses a Gaussian weighting function. Following [4] we test against the Gaussian the performance of two other weighting functions, drawn from the literature on robust statistics.

The second group of experiments tests the performance of the algorithm using the proximity matrix proposed in (3) against two other matrixes proposed in previous works [29,7], and a SIFT-based point matcher, based on the Euclidean distance between SIFTs, proposed by Lowe in [23], and used in [26] for measuring the SIFT performance.

For evaluating the performance of the three point matching methods used for this work we computed: (a) the total number of matches detected; (b) the number of correct matches; (c) the accuracy, defined as the ratio between number of correct matches and the total number of matches detected.

The data used are of different nature. We considered a stereo image sequence taken with a stereo system with relatively large baseline, and in particular we focused our experiments on input sequences for an immersive video-conferencing system [18]. Then we used short sequences with large variations with

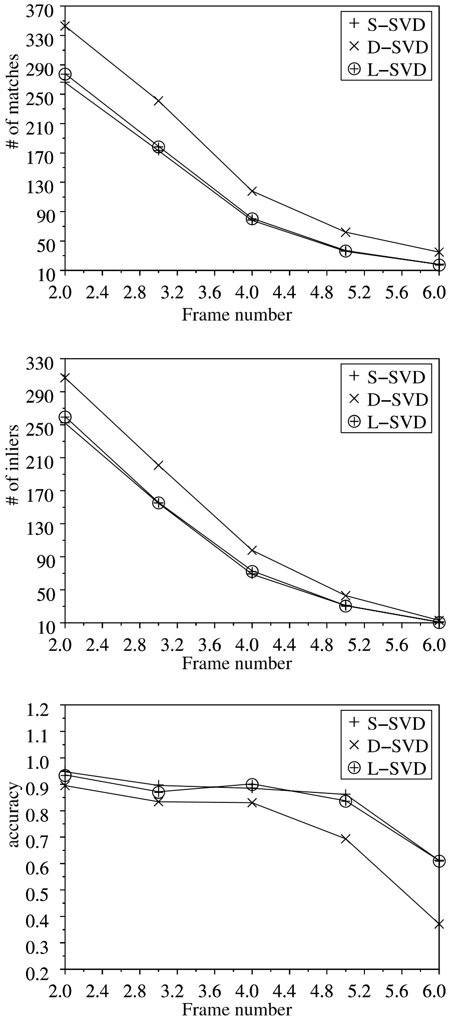

Fig. 8. Comparison of different weighting functions: results for the *Boat* sequence. The images are zoomed and rotated respect to the first frame. Matches are computed between the first frame and each other frame of the sequence. Top: total number of matches detected. Middle number of correct matches. Bottom: accuracy of the method.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

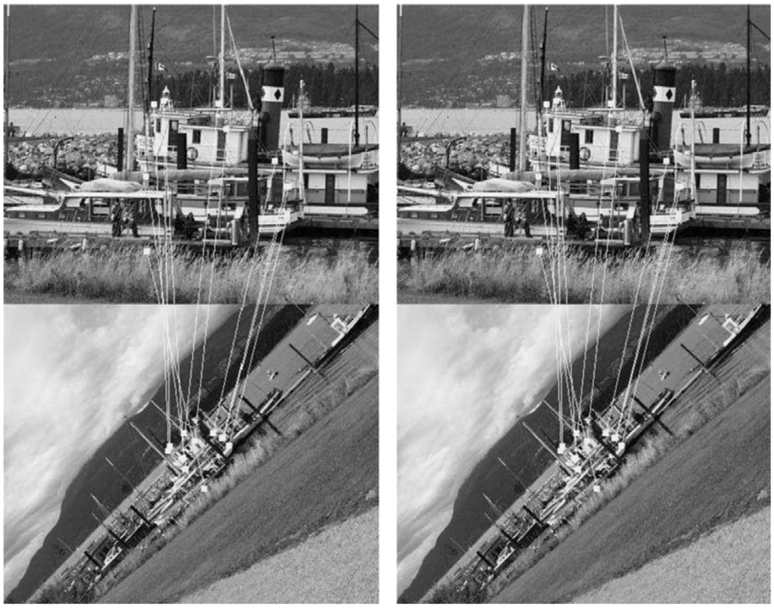

Fig. 9. Comparison of different weighting functions: results for *Boat* sequence. Correct matches between the left (top) and right (bottom) last. Left; S-SVD Right: L-SVD. The results for D-SVD are in Fig. 14.

respect to the first frame: the kind of variations considered are viewpoint changes and zoom plus rotation.[1] Some of these last sequences were used in [26]. The experiments performed on the video sequences compare the first frame of the sequence with all the others. Sample frames from the sequence used are shown in Fig. 4.

The method used for determining the correct matches depends on what geometric information on the camera geometry is available. For sets of data consisting of fixed camera sequences or sequences of planar scenes for which the homographies between the different views were available, we say that a pair of corresponding points (p,p') is a correct match if

$$\|p' - Hp\| < 5$$

where H is the homography between the two images. For the stereo-sequence with a fixed baseline the correspondence were computed between images of each stereo frame. In this case, because the scene is not planar, we compute the fundamental matrix F from the calibrated camera matrixes, and a pair of corresponding points $(p\ p')$ is a correct match if

$$(d(p', Fp) + d(p, F^t p'))/2 < 5$$

where $d(p',Fp)$ is the distance between point p' and the epipolar line corresponding to point p [41].

For all the experiments we set the parameter σ to 1000.

5.2.1. Comparison of different weighting functions

The weighting function models the probability of the similarity between the feature points. In previous works it was used the Gaussian weighting function. The reason for trying functions different from the Gaussian is that the distance between feature descriptors of corresponding points increases with the baseline. In this case a function with more prominent tails than the Gaussian can give the chance to detect some more matches. This, as we will see, at the price of a sometimes lower accuracy.

In this section we considered a small sample of different weighting functions borrowed from the literature on robust statics, in particular from the literature on M-estimators [15]. The comparative evaluation on the performance of different matching methods, whose description is given in 4, are based on the following different weighting functions:

- *S-SVD*: a Gaussian weighting function as it has been used all along the paper;
- *D-SVD*:a double-exponential weighting function

$$G_{ij} = e^{(-|r_{ij}/\|\sigma|)} \qquad (4)$$

[1] Sequences available from http://www.robots.ox.ac.uk/~vgg/research/affine/index.html

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

- *L-SVD*: a Lorentzian weighting function, defined as

$$G_{ij} = \frac{1}{1 + \frac{1}{2}\frac{r_{ij}^2}{\sigma^2}} \tag{5}$$

The different weighting functions are shown in Fig. 5. We choose

In Figs. 6 and 7 we show the results for the videoconferencing stereo sequence. We see that in terms of number of matches and correct matches the double-exponential function returns the best results, while the Gaussian and the Lorentzian have similar performance. These last two report an average accuracy of 0.6. The accuracy returned by the double-exponential is lower, but on average above 0.5, that means that at most 50% of the matches detected are wrong matches, and this is the largest amount of wrong matches that standard robust statistics tools can tolerate.

The results for the *Boat* sequence are shown in Figs. 8 and 9. Even in this case the D-SVD returns the highest number of correct matches, and, except for the last frame, the accuracy reported is above

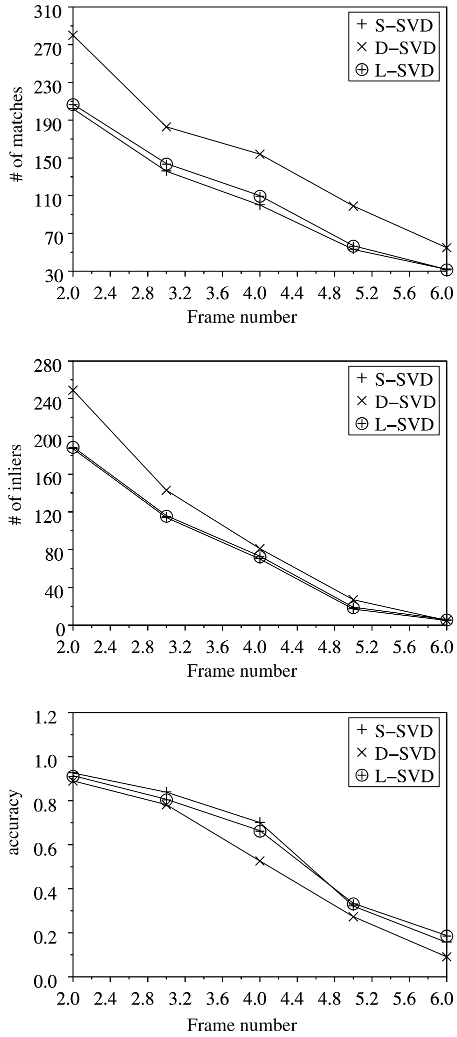

Fig. 10. Comparison of different weighting functions: results for the *Graf* sequence. The images present a change in the view-point respect to the first frame. Top: total number of matches detected. Middle: number of correct matches. Bottom: accuracy of the method.

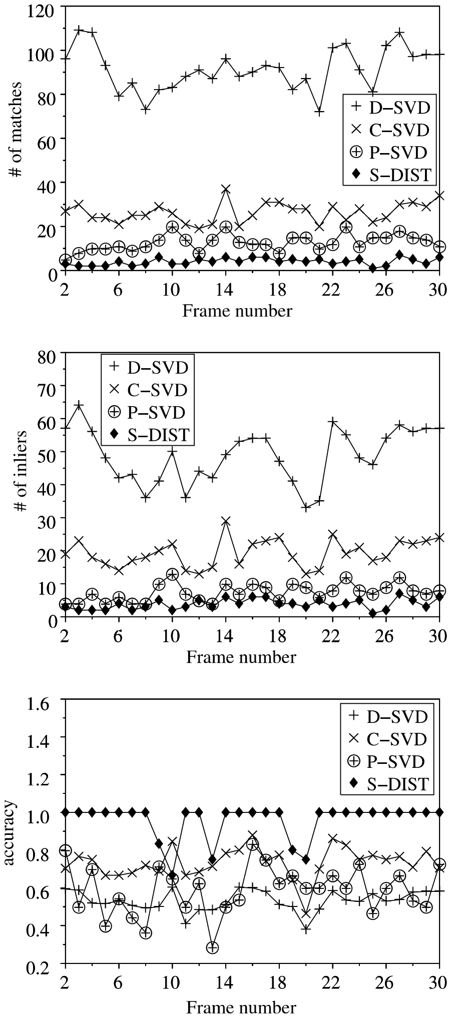

Fig. 11. Comparison with other algorithms: results for the 30-frames stereo sequence. The baseline is fixed for all the stereo pairs, and the correspondences are computed for each stereo frame of the sequence. Top: total number of matches detected. Middle: number of correct matches. Bottom: accuracy of the method.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

0.7. The other two functions reported an accuracy always well above 0.5.

For the *Graf* sequence the results are similar to what seen in the previous section (see Fig. 10). The double-exponential still returns the largest number of correct matches, but again the performance drops for the last two frames of the sequence when the change in the point of view is too large.

We can conclude this evaluation of the weighting functions saying that the double-exponential performs slightly better than the other two functions considered, but it does not seem that the use of any of these function dramatically changes the performance of the algorithm.

The double-exponential weighting function will be used in the following analysis.

5.2.2. Comparison with other matching algorithms

The comparative evaluation on the performance of different matching methods considers the following techniques:

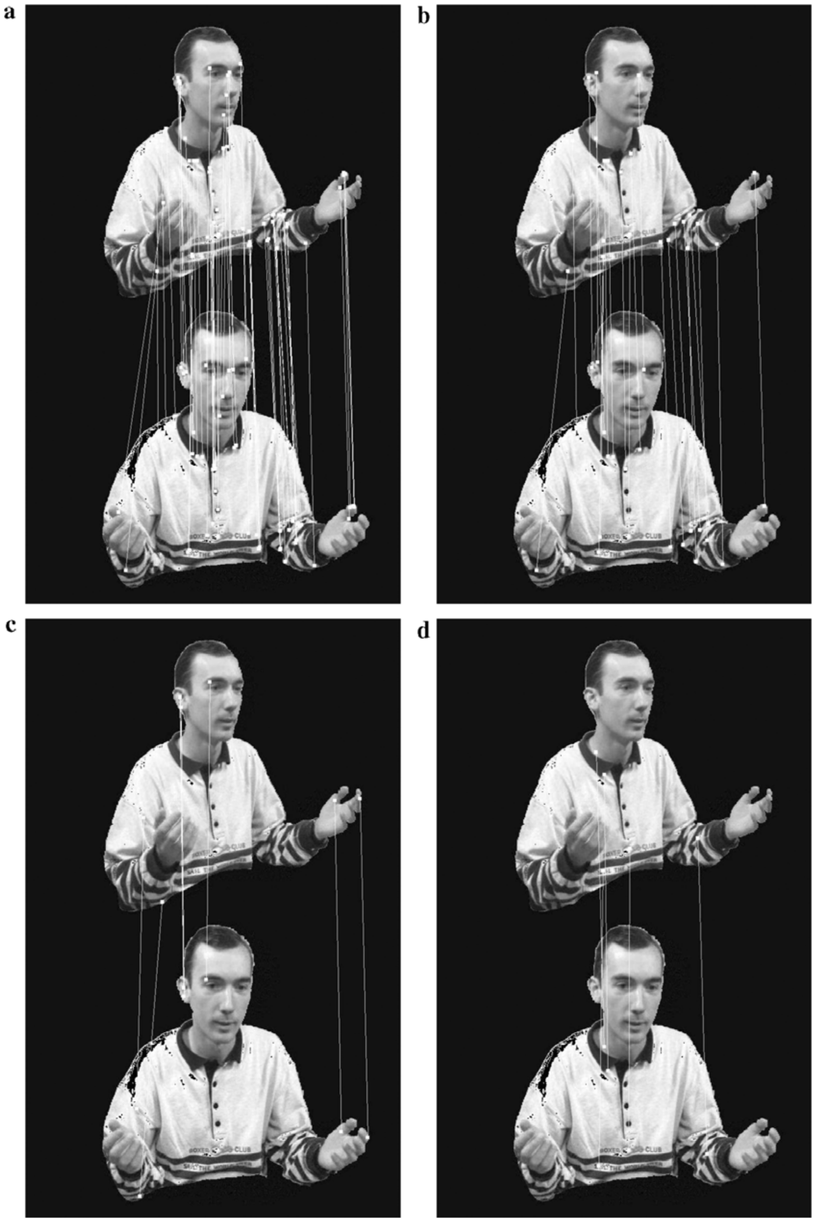

Fig. 12. Comparison with other algorithms: results for the 30-frames stereo sequence. Correct matches between the left (top) and right (bottom) 27th frames. (a) D-SVD. (b) C-SVD. (c) P-SVD. (d) S-DIST.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

- *D-SVD*: point matches are established following the SVD-matching algorithm of Section 3 with the proximity matrix **G** given in (4);
- *C-SVD*: point matches are established following the algorithm discussed in [7]

$$C_{ij} = \sum_t \frac{(S^i_t - \mathrm{mean}(S^i))(S^j_t - \mathrm{mean}(S^j))}{\mathrm{stdv}(S^i)\mathrm{stdv}(S^j)}$$

where S^i and S^j are the SIFT descriptors;

- *P-SVD*: point matches are determined as for C-SVD but with

$$c_{ij} = \sum_t \frac{(I^i_t - \mathrm{mean}(I^i))(I^j_t - \mathrm{mean}(I^j))}{\mathrm{stdv}(I^i)\mathrm{stdv}(I^j)}$$

where I^i and I^j are the two grey-levels neighbour;

- *S-DIST*: point matches are established following the method proposed in [23], that is two features *i* and *j* matches if

$$d_{ij} = \min(D_i) < 0.6 \min(D_i - \{d_{ij}\})$$

and

$$d_{ji} = \min(D_j) < 0.6 \min(D_j - \{d_{ji}\})$$

where $D_i = \{d_{ih} = \|S^i - S^h\|\}$.

In Figs. 11 and 12 we show the results for the video-conferencing stereo sequence. The S-SVD returns the largest number of matches and of correct matches (an average of 50 and 40, respectively, for each stereo frame) with respect to the other three: the C-SVD presents an average of 30 and 20 per stereo frame, while the values returned by the other two methods are much lower.

S-DIST returns the highest accuracy (almost always 1), but a very small number of matches. The accuracy obtained with D-SVD and C-SVD is slightly lower (ranging from 0.7 to 0.5) but it is high enough to use standard robust statistics tools for identifying and discarding wrong matches. As for P-SVD we notice that accuracy drops down to 0.4 that is too low for trating outliers with robust statistics.

The results shown in Figs. 13 and 14 are relative to a six frames sequence where the fixed camera zooms and rotates around the optical centre. In this case D-SVD is still giving the larger amount of correct matches. The number of matches goes down sensibly, because of the large zoom effect between the first and the last frame, so that the points detected at a finer scale in the first frame cannot be matched. The C-SVD still has acceptable performance while the other two methods perform poorly on this sequence. In particular P-SVD can only find matches between the first two frames. This is because this method uses correlation between image patches, that are very sensitive to rotation and scale changes.

The performance of the algorithms starts to go down with severe changes in the view-point, as shown in Figs. 15 and 16. In fact for the last 2 frames the amount of matches and the accuracy obtained are too low. The results returned by the S-DIST algorithm, that has been designed for the SIFT descriptor, are even worse, implying that the

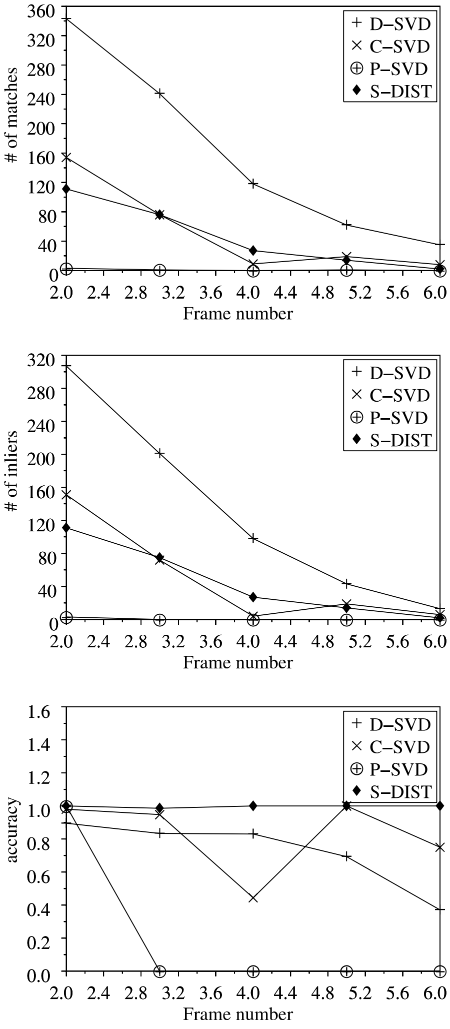

Fig. 13. Comparison with other algorithms: results for the *Boat* sequence. The images are zoomed and rotated respect to the first frame. Matches computed between the first frame and each other frame in the sequence. Top: total number of matches detected. Middle: number of correct matches. Bottom: accuracy of the method.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

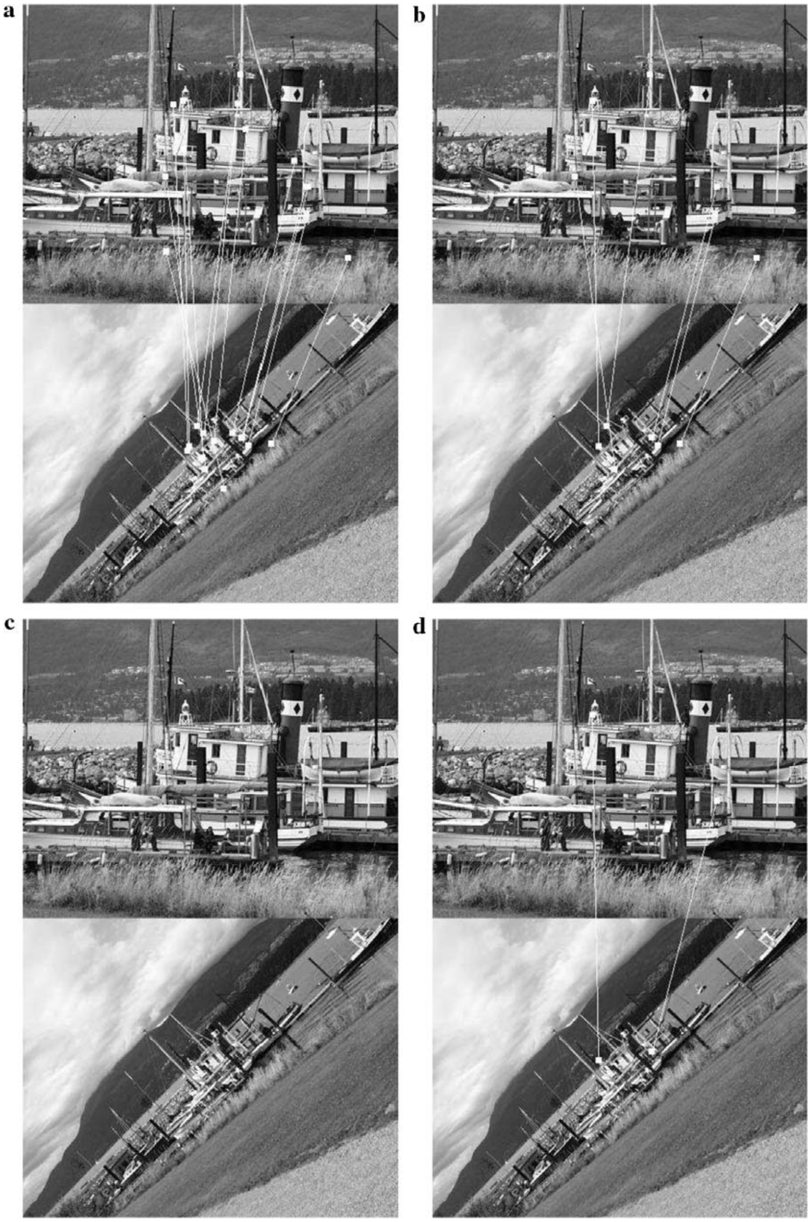

Fig. 14. Comparison with other algorithms: results for the *Boat* sequence. Correct matches between the left (top) and right (bottom) last frames. (a) D-SVD. (b) C-SVD. (c) P-SVD. (d) S-DIST.

descriptor cannot cope with too large viewpoint changes. Similar results have been reported in [26] for several descriptors.

In conclusion we can state that the use of the SIFT descriptors in combination with a SVD-matching algorithm improves the performance with respect to older versions of the algorithm, as already shown in [7]. Moreover the experiments reported in this paper show that replacing the distance between feature points with the distance between point descriptors in the weighting function used to build the proximity matrix gives better results when large changes in the scene occur. This is particularly noticeable in the case of severe zoom/rotation changes. However the performance are still not satisfactory for the case of too large viewpoint change.

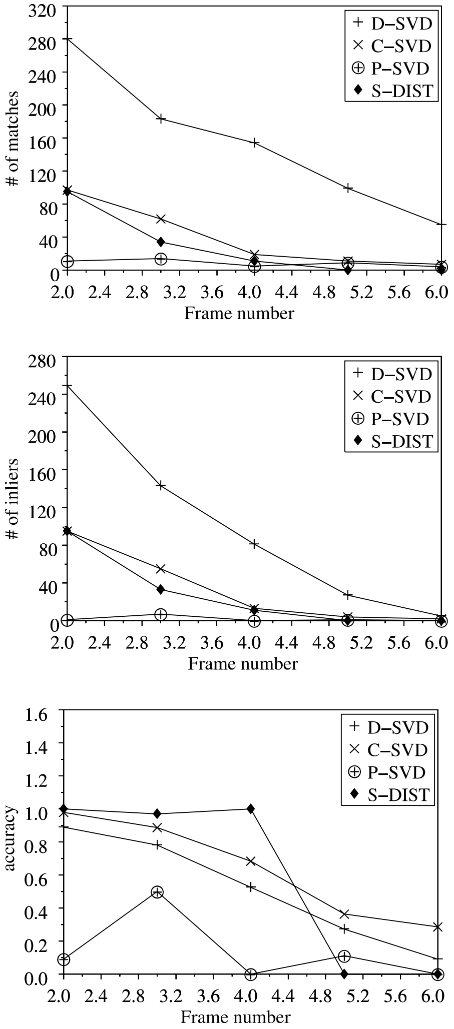

Fig. 15. Comparison with other algorithms: results for the *Graf* sequence. The images present a change in the view-point respect to the first frame. Top: total number of matches detected. Middle: number of correct matches. Bottom: accuracy of the method.

6. Conclusions

This paper presented a method for determining the correspondences between sparse feature points in images of the same scene based on the SVD-matching paradigm, that has been used by different authors in the past, and on a state-of-the-art keypoint descriptor, namely SIFT.

We showed that including SIFT point descriptors in the SVD-matching improves the performance with respect to the past versions of this algorithm. In particular it returned good results for scale changes, severe zoom and image plane rotations, and large view-point variations. The current version still does not cope with wide-baslines. These conclusions are supported by an extensive experimental evaluation, on different typologies of image data.

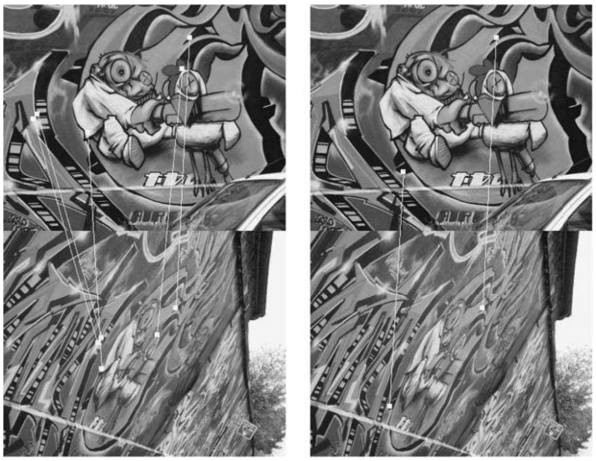

Fig. 16. Comparison with other algorithms: results for the *Graf* sequence. Correct matches between the left (top) and right (bottom) last frames. (a) D-SVD. (b) C-SVD. P-SVD. and S-DIST did not return any correct match.

As for many spectral methods, the SVD-matching algorithm is based on the choice of an appropriate weighting function for building a proximity matrix. The previous SVD-matching algorithms were using a Gaussian function. We compared its performance against other functions borrowed from the robust statistics literature: the Lorentzian and the double-exponential function. The results obtained suggest that the choice of this latter function can somewhat improve the quality of the results, but it does not appear to be a crucial issue.

Acknowledgments

The authors thank Maurizio Pilu for useful discussions.

The video-conferencing sequences used in this work have been provided by the EU Framework-V Grant (IST-1999-10044) VIRTUE (VIRtual Team User Environment).

The software used for detecting the keypoints is due to K. Mikolajczyk and is available for download from http://www.robots.ox.ac.uk/vgg/research/affine/index.html.

This work is partially supported by European FP6 NoE aim@shape grant 506766, the FIRB Project ASTA RBAU01877P.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

References

[1] N. Allezard, M. Dhome, F. Jurie, Recognition of 3D textured objects by mixing view-based and model based representations, in: Proceedings of ICPR, 2000.

[2] A. Baumberg, Reliable feature matching across widely separated views, in: Proceedings of CVPR, 2000, pp.774–781.

[3] M. Brown, D. Lowe, Invariant features from interesting point groups, in: Proceedings of BMVC, 2002, pp. 656–665.

[4] M. Carcassoni, E.R. Hancock, Spectral correspondence for point pattern matching, Pattern Recognition 36 (2003) 193–204.

[5] F.R. Chung, Spectral Graph Theory, American Mathematical Society vol,92 (1997).

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

[7] E. Delponte, F. Isgrò, F. Odone, A. Verri. SVD-matching using sift features. in: E. Trucco, M. Chantler, (Eds), Proceedings of the of the International Conference on Vision, Video and Graphics, pp 125–132, Edinburgh, UK, 2005, pp. 125–132.

[8] R. Deriche, Z. Zhang, Q.T. Luong, O. Faugeras, Robust recovery of the epipolar geometry from an uncalibrated stereo rig. in: J.O. Eklundh, (Eds), Proceedings of ECCV, 1994, pp. 567–576.

[9] S. Edelman, N. Ingrator, T. Poggio, Complex cells and object recognition. 1997. Unpublished manuscript. Cogprints.

[10] V. Ferrari, T. Tuytelaars, L.V. Gool, Wide-baseline muliple-view correspondences. in: Proceedings of CVPR, 2003.

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

[12] M. Galun, E. Sharon, R. Basri, A. Brandt, Texture segmentation by multiscale aggregation of filter responses and shape elements, in: Proceedings of ICCV, 2003, pp. 716–723.

[13] N. Georgis, M. Petrou, J. Kittler, On the correspondence problem for wide angular separation of non-coplanar points, Image and Vision Computing 16 (1998).

[14] G.H. Golub, C.F.V. Loan, Matrix Computations, John Hopkins University Press, 1983.

[15] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, 1986.

[16] C. Harris, Geometry from visual motion. in: A. Blake, A. Yuille (Eds), Active Vision, MIT Press, 1992.

[17] C. Harris, M. Stephens, A combined corner and edge detector, in: Proceedings of the 4th Alvey Vision Conference, 1988, pp. 147–151.

[18] F. Isgrò, E. Trucco, P. Kauff, O. Schreer, Three-dimensional image processing in the future of immersive media, IEEE Transactions on Circuits and Systems for Video Technology 14 (3) (2004) 288–303.

[19] T. Lindeberg, Principles for automatic scale selection, Technical Report ISRN KTH/NA/P-98/14-SE, CVAP Department of numerical analysis and computing science KTH, S-100 44 Stockholm, Sweden, August 1998.

[20] T. Lindeberg, Scale-space theory: a basic tool for analysing structures at different scales, Journal of Applied Statistics 21 (2) (1994) 224–270.

[21] M. Lourakis, S. Tzurbakis, A. Argyros, S. Orphanoudakis, Feature transfer and matching in disparate stereo views through the use of plane homographies, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2) (2003).

[22] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[23] David G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of ICCV, Corfu, Greece, 1999, pp. 1150–1157

[24] Xiaoye Lu, R. Manduchi, wide-basline feature matching using the cross-epipolar ordering constraint, in: Proceedings of CVPR, 2004, pp. 16–23.

[25] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceedings of BMVC, 2002, pp. 384–393.

[26] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in: Proceedings of CVPR, 2003, pp. 257–263.

[27] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision 60 (1) (2004) 63–86.

[28] H. Moravec, Rover visual obstacle avoidance, in: Proc. of the 7th Intern, Joint Conference on Artificial Intelligence, 1981, pp. 785–790.

[29] M. Pilu, A direct method for stereo correspondence based on singular value decomposition, in: Proceedings of CVPR, Puerto Rico, 1997, pp. 261–266.

[30] P. Pritchett, A. Zisserman, Wide baseline stereo matching, in: Proceedings of ICCV, 1998, pp. 754–760.

[31] A. Rosenfeld, G. van der Brug, Coarse-fine template matching, IEEE Transactions on Systems, Man and Cybernetics 7 (1977) 104–107.

[32] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (5) (1997) 530–534.

[33] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, International Journal of Computer Vision 37 (2) (2000) 151–172.

[34] C. Schmid, A. Zissermann, Automatic line matching across views. in: Proceedings of CVPR, (1997)666–671.

[35] S. Sclaroff, A.P. Pentland, Modal matching for correspondence and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (6) (1995) 545–561.

[36] G. Scott, H. Longuet-Higgins, An algorithm for associating the features of two images, Proceedings of Royal Society London B244 (1991) 21–26.

[37] L.S. Shapiro, J.M. Brady, Feature-based correspondence – an eigenvector approach, Image Vision Computing 10 (1992).

[38] J. Shi, C. Tomasi, Good features to track. in: Proceedings of CVPR, 1994, pp. 593–600,

[39] A. Shokoufandeh, I. Marsic, S.J. Dickinson, View-based object recognition using saliency maps, Image and Vision Computing 17 (1999) 445–460.

[40] D. Tell, S. Carlsson, Combining appearance and topology for wide baseline matching. in: Proceedings of ECCV, 2002, pp. 68–81.

[41] E. Trucco, A. Verri, Introductory Techniques for 3-D Computer Vision, Prentice-Hall, 1998.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.

[42] T. Tuytelaars, L.V. Gool. Wide baseline stereo matching based on local, affinely invariant regions, in: Proceedings of BMVC, 2000, pp. 412–425.

[43] S. Umeyama, An eigen decomposition approach to weighted graph matching problems, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (1988).

[44] G. van der Brug, A. Rosenfeld, Two-stage template matching, IEEE Transactions on Computers 26 (4) (1977) 384–393.

[45] Y. Weiss, Segmentation using eigenvectors: a unifying view. In Proceedings of ICCV, pages 975–982, 1999.

[46] R. Zabih, J. Woodfill. Non-parametric local transforms for computing visual correspodence. In Proceedings of ECCV, pages 151–158, 1994.

[47] Z. Zhang, R. Deriche, O. Faugeras, Q.T. Luong, A robust techique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, Artificial Intelligence 78 (1995) 87–119.

Please cite this article as: Elisabetta Delponte et al., SVD-matching using SIFT features, Graphical Models (2006), doi:10.1016/j.gmod.2006.07.002.