Large Baseline Matching of Scale Invariant Features

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

INFM - DISI, Università di Genova, Genova, Italy

Abstract. The problem of feature points matching between pair of views of the scene is one of the key problems in computer vision, because of the number of applications. In this paper we discuss an alternative version of an SVD matching algorithm earlier proposed in the literature. In the version proposed the original algorithm has been modified for coping with large baselines. The claim of improved performances for larger baselines is supported by experimental evidence.

1 Introduction

A number of computer vision tasks require that a set of point correspondence is established between image pairs, just to name a few: 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.

We are currently interested in what we call *large* baseline matching, i.e. for images taken with the same camera and from not close view points. For the present work we do not aim to give a solution for the wide-baseline problem, when the images can look too much different. Our interest to this problem come from a 3D data acquisition system we are developing, for which we are dealing with the problem of registering different measures of the same object coming from different poses. Even if methods for 3D data registration do exist [1], they do need a good initialisation, that might be provided by point matched in the images, providing that an appropriate calibration between an image and a 3D point cloud is given.

We aim at obtaining a system to compute sparse correspondences between image pairs and use them to initialise a 3D registration procedure. From the feature matching standpoint our goal is to devise a procedure that allows us to obtain a reasonably high number of accurate matches from image pairs acquired by a large baseline system.

The problem of matching points between images is covered by a very rich literature. We do consider here the case when the epipolar geometry is not known, and then the corresponding point can be anywhere in the image. A classical approach, for the case of short baseline, is the one presented in [2], that adopts a standard correlation step, followed by a relaxation step. For the case of *wide* baseline we mention the work by Pritchett [3], Baumberg [4], Tuytelaars [5].

Recently a simple constraints that can reduce the computational complexity for wide baseline matching, for the only case of parallel epipolar lines, has been proposed in [6].

We propose in this paper a simple modification of the SVD-based matching algorithm proposed in [7]. The core of the modification is to associate to each feature point a meaningful local descriptor to deal with larger baseline. In our method we extract from each image interesting points using difference of Gaussians, and represent them with the SIFT descriptor [8], making them tolerant to change and scale orientation. Finally our algorithm computes correspondences between the two sets of feature points with a mixed geometric and appearance based approach, as in [7]. In this last paper it is shown that a reasonably good solution to the matching problem can be achieved by an appropriate transformation of a correspondence strength matrix. The method, that builds on the top of a previous work by Scott and Longuet-Higgins [9], obtains the transformation of the strength matrix by singular value decomposition.

We report promising results, discussing how the matching behaves while the baseline grows. A comparison with the original work shows how SIFT features make the system more tolerant to the effects of a larger baseline. We also report a comparison to the matching criterion for SIFT key-points proposed by Lowe, showing that our approach is a better compromise between accurate results and a high number of correspondences.

The paper is organised as follows. In the next Section we review the SVD matching algorithm. In Section 3 the scale invariants features and their associated descriptor are briefly reviewed. Our modified SVD matching is described in Section 4, and experimental results are shown and discussed in Section 5. Section 6 is left to the final remarks.

2 SVD Matching

In this section we briefly review the matching algorithm proposed in [7]. As mentioned before this algorithm builds on the top of [9], and adapts it to deal with pixel correspondences. In the original work described in [9] it is shown that, in spite of the well-known combinatorics complexity of feature correspondence, a reasonably good solution can be achieved through the singular value decomposition of an appropriate correspondence strength matrix. In [7] this matrix is modified in order to consider intensity values together with spatial position of the feature points.

Let A and B be two images, from which we extracted m and n features respectively (A_i , $i = 1, \dots, m$, and B_j , $j = 1, \dots, n$), each of them represented by a simple image patch of size $w \times w$. The goal of the algorithm is to put two subsets of the two sets of features in a one-to-one correspondence.

The algorithm consists of three steps:

1. Build a correspondence matrix $\mathbf{G}$ that models both geometric proximity and similarity; each entry G_{ij} is computed as

$$G_{ij} = \frac{C_{ij} + 1}{2} e^{-r_{ij}^2 / 2\sigma^2}. \quad (1)$$

$r_{ij} = \|A_i - B_j\|$ is the Euclidean distance between the two features in the image plane, and C_{ij} is the normalised correlation between the two image neighbourhood. 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 $\mathbf{G}$ range from 0 to 1, with higher values for more correlated features.

2. Compute the Singular Value Decomposition for $\mathbf{G}$: $\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{U}^\top$.
3. Compute a new correspondence matrix $\mathbf{P}$ by converting diagonal matrix $\mathbf{D}$ to a diagonal matrix $\mathbf{E}$ where each element D_{ii} is replaced with a 1: $\mathbf{P} = \mathbf{V}\mathbf{E}\mathbf{U}^\top$. It is shown in [9] that $\mathbf{P}$ carries similar information of $\mathbf{G}$, with the interesting property of enhancing good pairings.

In [7] experimental evidence is given that the proposed algorithm performs well on short baseline stereo pairs. In fact the performance fall as 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.

3 Features Detection in Scale-Space

Scale invariant features transform key-points (SIFT) were first proposed in [10] and attracted the attention of the computer vision community for their tolerance to scale changes, illumination variations, and image rotations. These features are also claimed robust to affine distortion, change of viewpoints and additive noise.

A recent comparative study presented in [11] shows that SIFT descriptors are more stable than other state of the art interest point descriptors. Therefore we decided to focus our attention on this feature descriptor, associated to the DoG key-point. In the remainder of this section we will briefly introduce the key-points and a possible description.

3.1 Scale-Space and Interest Points Detection in Images

When approaching to computer vision one of the first remarks is that every object in an image assumes a different significance if observed at a different scale. It has been demonstrated [12] that *scale-space* is a good framework to handle objects in images at different scale. Indeed, scale-space is a representation of the image which is seen at different resolution levels while its fine-scale structures are deleted.

An important aspect of the scale-space is that the suppression of details is not a random process: scale-space keeps the meaningful information obtained at the different scales.

3.2 Scale Invariant Description of Interest Points

The process of building SIFTs [8] is heavily inspired by the scale-space framework, but it keeps all the information related to the different levels of resolution.

The process can be sketched in two phases: the first is key-points detection in scale-space pyramid and the second is key-points description using the image gradient at the right level of resolution. Key-points are detected in a particular pyramidal structure that is built as follows:

1. The original image I is convolved with a Gaussian function with $\sigma = \sqrt{2}$. The result I_{σ_1} is convolved again with the same Gaussian to obtain a double-filtered image I_{σ_2} .
2. The difference between these two images is called difference-of-Gaussian (DoG): $D_{12} = I_{\sigma_1} - I_{\sigma_2}$. This will be the first level of the DoG pyramid.
3. The image is sub-sampled and the Gaussian convolution is repeated to obtain the other levels of the pyramid of DoG.

Once the pyramid of DoGs is completed, maxima and minima are located. The feature detection phase ends with a cleaning procedure for discarding low contrast features, for filtering out edges.

Regions detected by DoG are mainly blob-like structures. There are no significant signal changes in the centre of the blob and therefore the Gaussian filter-based descriptors perform better in larger point neighbourhood [11]. Once the keypoints are accurately located in the scale-space, a principal direction is assigned to each of them. This orientation is attributed in order to achieve invariance with respect to rotations.

Following [8], we associate each key-point to a descriptor, computed as a composition of direction histograms in the neighbouring regions of the scale-space, shifted according to the dominant orientation of the feature.

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 introduction the SVD matching presented in [7] 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. It is now well known that image neighbour grey level values is a descriptor too sensitive to changes in the view-point, and more robust descriptor have been introduced so far (see, for instance, [13,5,14]).

Results of a comparative study, performed on a set of planar scenes, of the performance of various features descriptors have been reported in [11], where it is shown that the SIFT descriptor is better than the other descriptors with respect 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 not stable performance, depending on the accuracy of the point, that depends strongly on the kind of transformation considered.

The considerations above suggested the use of a SIFT descriptor, instead of grey levels. We left the matrix $\mathbf{G}$ in equation (1) unchanged in its form, but C_{ij} is now the cross-correlation between SIFT descriptors. As it will be shown in

the next Section, 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 the SIFT distance used for the experiments reported in [11]. We do plan to experiment with different SIFT distances in the SVD matching, which might require to modify the form of the $\mathbf{G}$ matrix.

5 Experimental Results

In this section we report experiments carried on short image sequences where the camera was moving around a *complex* indoor scene (i.e., with several objects), increasing the baseline with respect the camera pose for the first frame in the sequence. For reason of space we show here only results on a sequence of 13 frames.

For each frame in the sequence we extracted a set of interest points, using the DoG points detector described in Section 3, that proved invariant to rotation and scale changes [10,11]. We remind here that the points detected are local scale-space extrema of the difference of Gaussians. The size of the support region, the area used for computing the associated descriptor, is determined from the selected scale.

We compare the performance of the SIFT based SVD matching, henceforth S-SVD, against the performance of the correlation based one (C-SVD), and a SIFT based point matcher proposed by Lowe in [10], and used in [11] for measuring the SIFT performance, which uses the Euclidean distance between SIFTs. We will address to this last matching method as S-DIST. More formally the correspondence are established as

- **S-SVD**: point matches are established following the algorithm of Section 2

$$C_{ij} = \sum_t \frac{(S_t^i - \text{mean}(S^i))(S_t^j - \text{mean}(S^j))}{\text{stdv}(S^i)\text{stdv}(S^j)}$$

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

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

$$c_{ij} = \sum_t \frac{(I_t^i - \text{mean}(I^i))(I_t^j - \text{mean}(I^j))}{\text{stdv}(I^i)\text{stdv}(I^j)}$$

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

- **S-DIST**: two features i and j matches if

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

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

For measuring the goodness of a point match we computed the fundamental matrix [15] between the first frame and all the successive frames for each image

Fig. 1. Results for a 13-frames image sequence. Left: total number of matches detected. Centre: number of correct matches. Right: accuracy of the method.

Fig. 2. Results for a 13-frames image sequence. Correct matches between first (top) and sixth (bottom) frame. Left: S-SVD. Centre: C-SVD. Right: S-DIST.

sequence, using sets of point correspondences established by hand. Having the fundamental matrix $\mathbf{F}$ allows to verify if a match is correct, by using the point to epipolar line distances. We say that a match (p_i, p_j) is correct if the

$$d(p_i, \mathbf{F}p_j)^2 + d(p_j, \mathbf{F}^t p_i)^2 < 4.$$

We would like to point out that this measure is not bullet-proof, as it can happen that wrongly matched points lie on corresponding epipolar lines. In our experience when this situation occurs, it is no more than one or two points. In Figure 3 we show stereo pair for which this happened in the case of the S-DIST matcher.

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 plot of these three values, relative to the 13-frames sequence for which we show results, can be found in Figure 1.

Overall we can say that S-SVD outperforms C-SVD in all the cases. In general S-SVD returns the largest number of correct point matches, and total number

Fig. 3. Results for a 13-frames image sequence. Correct matches between first (top) and eightieth (bottom) frame. Left: S-SVD. Centre: C-SVD. Right: S-DIST. One of the matches in the right image is wrong but the two points lie on corresponding epipolar lines, therefore is wrongly passed as correct.

of point matches. In terms of accuracy the best one is S-DIST, but S-SVD has an accuracy of more than 0.5 for almost half the length of each sequence, that makes the quality of the matches good enough for any state of the art robust estimator [16]. On the other hand, the number of correct matches detected by the S-DIST for the cases when the accuracy of the S-SVD is below 0.5, is very often too small (2-6) for being useful for any task.

The quality of the matches goes down for all the three methods as the baseline starts to be too large, meaning that none of the methods can be feasible for wide-baseline matching, and some more work needs to be done in attempting to make the S-SVD algorithm more robust to the baseline variation.

In Figure 2 and 3 we show results of the matching between the first frame of the sequence and the sixth and the eightieth frame respectively.

6 Conclusions

In this paper we described an improved version of the SVD matching presented in [7] that is capable to deal with stereo pairs with reasonably large baseline. The improvement is obtained by using a more robust feature descriptor (SIFT) than the one used in the former version of the algorithm. Experimental evidence shows the better performance of the proposed version with respect the original one, and with respect a standard SIFT based point matcher.

More work is still necessary in trying to make the algorithm feasible for wide baseline matching when the images can look too much different. In our view what should be tried is: the use of a different interest point detector as the *improved* Harris point detector discussed in [11], and the use of SIFT distance measures different from the cross-correlation used in the current version of the S-SVD.

Acknowledgements

The authors thank Maurizio Pilu for useful discussions. This work is partially supported by European FP6 NoE AIM@SHAPE grant 506766, the INFM Advanced Research Project MAIA, the FIRB Project ASTA RBAU01877P.

References

1. Besl, P., McKay, N.: A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14** (1992) 239–256
2. Deriche, R., Zhang, Z., Luong, Q.T., Faugeras, O.: Robust recovery of the epipolar geometry from an uncalibrated stereo rig. In Eklundh, J.O., ed.: *Proceedings of the European Conference on Computer Vision*. Volume 800 of LNCS. (1994) 567–576
3. Pritchett, P., Zisserman, A.: Wide baseline stereo matching. In: *Proceedings of the International Conference on Computer Vision*. (1998) 754–760
4. Baumberg, A.: Reliable feature matching across widely separated views. In: *Proceedings of CVPR*. (2000) 774–781
5. Tuytelaars, T., Gool, L.V.: Wide baseline stereo matching based on local, affinely invariant regions. In: *Proceedings of the British Machine Vision Conference*. (200) 412–425
6. Lu, X., Manduchi, R.: Wide baseline feature matching using the cross-epipolar ordering constraint. In: *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*. Volume I. (2004) 16–23
7. Pilu, M.: A direct method for stereo correspondence based on singular value decomposition. In: *Proceedings of CVPR, Puerto Rico* (1997) 261–266
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60** (2004) 91–110
9. Scott, G., Longuet-Higgins, H.: An algorithm for associating the features of two patterns. *Proc. Royal Society London* **B244** (1991) 21–26
10. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proceedings of the ICCV, Corfú, Greece* (1999) 1150–1157
11. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: *Proceedings of CVPR*. Volume II. (2003) 257–263
12. Lindeberg, T.: Scale-space: A framework for handling image structures at multiple scales (1996)
13. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: *Proceedings of the European Conference on Computer Vision*. (1994) 151–158
14. Freeman, W., Adelson, E.: The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13** (1991) 891–906
15. Hartley, R.I., Zisserman, A.: *Multiple view geometry*. Cambridge University Press (2000)
16. Meer, P., Mintz, D., Rosenfeld, A., Kim, D.Y.: Robust regression methods for computer vision: a review. *International Journal of Computer Vision* **6** (1991) 59–70