A ROBUST METHOD FOR THE ESTIMATION OF RELIABLE WIDE BASELINE CORRESPONDENCES

Francesco Colletto, Marco Marcon, Augusto Sarti, Stefano Tubaro

Politecnico di Milano – Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci 32, Milano, Italy
colle1@tin.it; marcon/sarti/tubaro@elet.polimi.it

ABSTRACT

In this paper we present a complete method to retrieve reliable correspondences among *wide baseline* images, that is images of the same scene/object acquired from very different viewpoints. We propose a solution based on matching of affine covariant features, composed by the following four steps: interest region detection, normalization, description and matching. In our method we implemented improved versions of some techniques recently introduced in the literature: the *MSER* detector (*Maximally Stable Extremal Regions*) and *SIFT* and *RIFT* descriptors (*Scale / Rotation Invariant Feature Transform*). After a general introduction to the wide baseline problems and a summary of the recent state-of-the-art solutions, we illustrate the proposed method detailing the added improvements, then we present some experimental results obtained on wide baseline images.

Index Terms— Feature extraction, Image matching, Image region analysis, Robustness, Stereo vision

1. INTRODUCTION

The retrieval of correspondences among two or more shots of the same scene from different viewpoints is a fundamental step for several applications (such as camera reconstruction, object recognition, robot vision; see [1]). The solution of this problem was deeply analyzed for *short baseline* images, i.e. images where the viewpoints distance is relatively small with respect to the closest object distance. In this case the perspective distortions are limited and corresponding points can be easily found: in a typical short baseline method, key-points (usually corners) are detected in both images and a region around each of them is considered for matching; comparisons are then made between regions that occupy similar places in the considered images and putative matches are obtained according to a distance criterion of the neighbourhoods.

This approach fails if applied to wide baseline images, being unable to cope with consistent perspective distortions, rotations, occlusions and light change, that are typical wide baseline problems. In the last few years some solutions have been proposed, well summarized in [1]. This is the basic idea: unlike for the short baseline case, it is necessary to consider features whose shape and size are not fixed but adapted from the estimated local perspective transformations; each feature then will be normalized and compared with all the other. In practice, the detected regions of interest are small, so they can be assumed to belong to local planar 3D surfaces. Being perspective effects generally negligible on such a scale, the homography relating corresponding features can then be approximated with an affine transformation. The detected features have shape and size varying with local affinities and are called *affine covariant* regions [1].

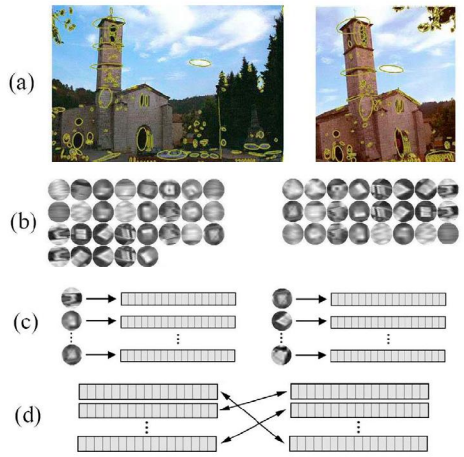


Fig. 1. Illustration of the four steps of the proposed method: detection(a), normalization(b), description(c), matching(d).

2. THE PROPOSED METHOD

Following the guidelines exposed in the recent literature, our proposed method is composed by the following four steps: feature detection, normalization, description and matching (see Figure 1). Several user-modifiable parameters are introduced for each step to allow a fine control of the entire process. In our implementation, as shown below, we chose restrictive default values, to retrieve a medium number of highly reliable correspondences with a low rate of false matches (outliers).

2.1. Detection

The detection step extracts, from each image, affine covariant regions representing the features to be matched. In our work we implemented a detector of *Maximally Stable Extremal Regions* (*MSER*) [2], adding some further improvements described below. We chose the *MSER* detector since it shows good performances and, compared in [1] to other detectors of the literature, it resulted in the most efficient and obtained the best test scores in case of large viewpoint and illumination changes, that are of major interest in our work.

A *MSER* is a connected region of pixels characterized by an almost constant intensity that can be well distinguished from its outer boundaries. The *MSER* detection algorithm is similar to the watershed one and operates applying increasing thresholds to the image luminance. At each threshold level I, the existing connected regions (called *Extremal Regions*, briefly *ER*) are evaluated (new *ER*s may appear, or pre-existing *ER*s may grow or merge). For each luminance level I the area $A(I)$ of each existing *ER* is stored; a stability analysis is then performed for each $A(I)$ function and the region boundaries are chosen accordingly to two different stability criteria (see Figure 2). The first one follows Matas et al. in [2] with some modifications: for each region the following function $R(I)$ is computed from $A(I)$:

$$R(I) = \frac{A(I+\Delta) - A(I-\Delta)}{A(I)} \qquad (1)$$

It can be considered a symmetric incremental ratio of A where Δ is a parameter to control selectivity: the local minima of $R(I)$ are considered as maximally stable values of I for that region. To improve the detection results, we introduced additional threshold parameters to be more selective: ΔI_{min} allows to discard regions whose global intensity variation is low, while A_{min} and A_{max} thresholds are introduced to discard regions with area lower than A_{min} (too small to represent a good feature pattern) or greater than A_{max} (too big for planarity approximation); finally, the $\Delta A_{min\%}$ threshold represents the minimum percentage of area increment between two consecutive *MSER*s derived from the same *ER*, to avoid the extraction of multiple similar features.

The second stability criterion proposed in this paper differs from the first one since it is independent from single local minima and is based on stability intervals, as illustrated below (Figure 2(c)). Given $A(I)$, the percentage of area increment function $\Delta A_{\%}(I)$ is computed as:

$$\Delta A_{\%}(I) = \frac{A(I) - A(I-1)}{A(I-1)} \qquad (2)$$

Stability intervals of $\Delta A_{\%}(I)$ are searched, i.e. luminance intervals $[I_0,\ I_1]$ where $\Delta A_{\%}(I)$ is lower than $\Delta A_{max\%}$; if the length of the interval (that is $I_1 - I_0$) is greater than the minimum value specified by the ΔI_{stab} parameter, then the mean value of the interval (i.e. $(I_1 - I_0)/2$) is assumed as maximally stable and determines a *MSER* contour. The parameters introduced above for the first criterion (A_{min}, ...) are also used to improve the results.

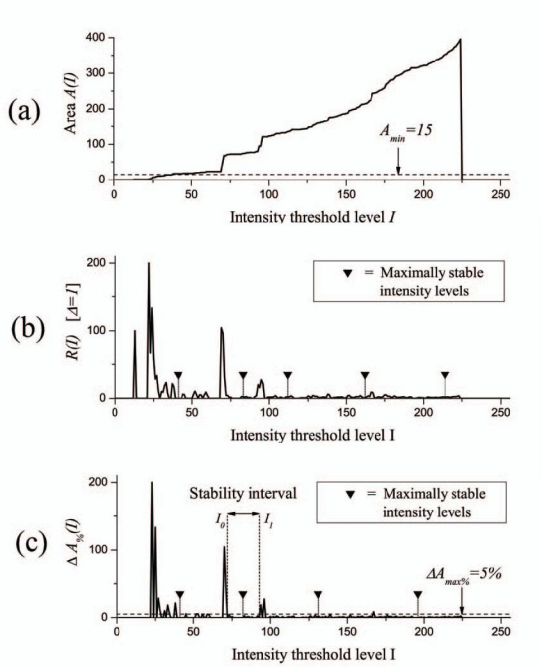


Fig. 2. A typical area function $A(I)$ of a single *Extremal Region* (a) and application of the implemented first (b) and second (c) stability criterion.

Experimentally the proposed stability criteria give similar performances and both of them result in most of cases superior to the original *MSER* algorithm, as showed in Section 3.

2.2. Normalization

In this step, the detected affine covariant regions (*MSER*s) are normalized to become *affine invariant*, so comparable between images. Firstly, elliptical regions are derived from the detected irregular-shaped *MSER*s through an ellipse fitting method, then each ellipse is enlarged by a factor S (default is $S = 2.5$) to cover a wider image area and therefore include additional neighborhood information. This rescaling allows to increment the descriptive power of each feature and lead to a noticeable gain in matching performances; however, as pointed out in [1], too high values of S can worsen the results, due to the higher risk of occlusions or non-planarities. A geometric and photometric normalization is then applied to the scaled elliptical regions, respectively an affine transformation from ellipse to circle (with bilinear interpolation of luminance values) and a normalization of pixels intensities from 0 to 1. The result are circular regions of normalized intensities.

After this step, corresponding circular regions extracted from different images should be almost identical, except for an unknown rotation factor: a final uniform alignment is required to achieve rotational invariance. This is done eval-

uating the dominant gradient orientation(s) (*DGO*) for each region and aligning it to the horizontal axis. The adopted method is similar to the one presented in [3], but in our implementation a larger orientation histogram (128 bins), smoothing and parabolic interpolation are used to obtain a more accurate estimation of the *DGO*s.

2.3. Description

The aim of the descriptor is to summarize the characteristic information of a normalized circular region in a numerical feature vector of small size (generally lower than 200 elements); a good descriptor should be distinctive and at the same time robust to geometric and photometric changes.

In our work we implemented two types of descriptors, *SIFT* [3] and *RIFT* [4] (*Scale / Rotation Invariant Feature Transform*, with some modifications. We chose the *SIFT* descriptor since it performs better in the comparative tests in [5].

A description vector is generated as follows: the input normalized circular region is divided in sub-regions, squares for *SIFT* descriptor and concentric rings of the same area for *RIFT*; for each sub-region the histogram of gradient orientations is computed, using the gradient magnitudes as weight (histograms have H bins, default is $H = 8$, i.e. one each 45 degrees); the output vector is then constructed concatenating the values of the histograms of the different sub-regions. *SIFT* and *RIFT* descriptors are characterized by good robustness and descriptive power; from our experiments (Section 3), *RIFT* resulted more efficient than *SIFT* being rotation invariant by design so not requiring the uniform alignment of the normalized regions, but the latter showed a more discriminative power leading to a larger number of correspondences.

2.4. Matching

In this final step the description vectors from different images are compared and a putative feature match is returned if the distance between compared vectors satisfies a particular criterion. In our work we implemented a *Nearest Neighbour Distance Ratio* (briefly *NNDR*) criterion, since in [5] it resulted the best choice if coupled with a *SIFT*-type descriptor.

Considering a description vector $\mathbf{u}$ from image U and the vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ from image V having respectively the minimum and the second smallest Euclidean distance from $\mathbf{u}$, the correspondence $\mathbf{u} \leftrightarrow \mathbf{v_1}$ is considered a valid match if the distances satisfy:

$$\frac{d(\mathbf{u}, \mathbf{v_1})}{d(\mathbf{u}, \mathbf{v_2})} \leq R_{min} \qquad (3)$$

The correspondence between $\mathbf{u}$ and its nearest neighbor $\mathbf{v_1}$ is then accepted if the minimum distance $d(\mathbf{u}, \mathbf{v_1})$ is significantly lower (depending on R_{min} parameter) than the distances between $\mathbf{u}$ and the other $\mathbf{v_k}$ vectors. The *NNDR* criterion is very selective and returns a very reliable set of correspondences; the default value for the R_{min} is 0.75, since we found it gives good results and greater values lead to greater percentages of false matches.

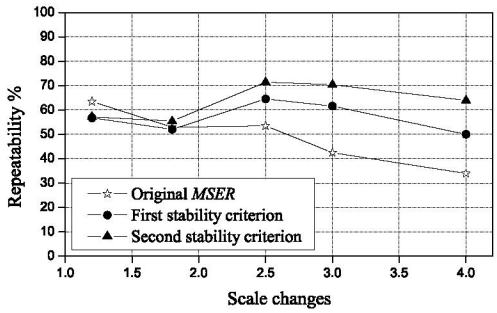


Fig. 3. Repeatabily scores for the *bark* reference test set of the original *MSER* algorithm and our improved version (for the two implemented stability criteria).

3. EXPERIMENTAL RESULTS

In our experiments, we tested the performances of the implemented *MSER*-type detector using the reference framework[1] introduced in [1]. The obtained scores are generally superior than the original *MSER* algorithm, for both the implemented stability criteria (in Section 2.1) and in some cases are noticeably improved, as showed in Figure 3.

In order to test the reliability of the retrieved wide baseline correspondences, we also executed experiments of epipolar geometry estimation: applying the presented method to every pair of images of the *Valbonne* set, the returned matches have been evaluated using a *RANSAC*-type robust algorithm. The obtained results are pretty good: using the default parameter values, a medium number of highly reliable matches is retrieved, even in cases of large viewpoint change and illumination change (some examples are shown in Figure 4). A mean number of 117 *SIFT* correspondences (52 for *RIFT*) is returned for consecutive images of the set (nearly short baseline cases), and it decreases while the baseline increments (for the extreme case, i.e. the first and the last image of the set, the matches found are 4, with one incorrect); the mean rate of false matches is about 9%. The *SIFT* descriptor generally returned a larger number of correspondences than *RIFT* (nearly two times greater on average), with similar percentages of outliers, but the latter required less computation time. Indicative execution time for a pair of images of the *Valbonne* set (512×768 pixels) is about 24 sec using *SIFT* and about 19 sec using *RIFT* (on a Athlon64 2.0GHz machine with 1Gb of RAM DDR400).

Finally, we compared our method with the complete original *SIFT* algorithm, developed by Lowe[2], composed by a multi-scale key-point detection (by difference-of-Gaussian) and the region description step illustrated above. In case of

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

[2] available at `http://www.cs.ubc.ca/~lowe/keypoints/`

short-medium baseline, this algorithm returned a larger number of matches than our method (even two times greater), maintaining a similar rate of outliers; however, as showed in Table 1, this approach fails if the viewpoint change is more consistent: in these cases our solution proved to be more robust, still leading to reliable matches.

	Orginal *SIFT*	Our method
Graffiti img 1–6	9 / all outliers	31 / 3 outliers (9.7%)
Wall img 1–6	2 / 1 outlier (50%)	16 / no outliers

Table 1. Number of matches and outliers for the original *SIFT* algorithm and our proposed method; high viewpoint change, structured and textured scene (*Graffiti*, Figure 3(c), and *Wall*).

4. CONCLUSIONS AND FUTURE WORKS

In this work we proposed a complete wide baseline matching method that improves some state-of-the-art ideas and allows us to successfully achieve reliable feature correspondences even in difficult cases. We focused on stereo images, but the matching step can be easily expanded to the case of multiple wide baseline images: after considering each pair of them, a more accurate set of multiple correspondences can be retrieved exploiting the more restrictive constraints arising from the multiple view geometry or the local affinities, allowing a guided search for further matches. Possible future expansions are combination of *MSER* with other detectors (*Hessian-affine* [1] for example), combination of *SIFT* with other descriptors, improvement of the *RIFT* descriptor, extension to the multiple view case and guided matching of new correspondences to allow a possible wide baseline 3D reconstruction.

5. REFERENCES

[1] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors", *IJCV*, vol. 65, no. 1/2, pp. 43–72, 2005.

[2] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions", in *Proc. BMVC*, 2002, pp. 384–393.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints", *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

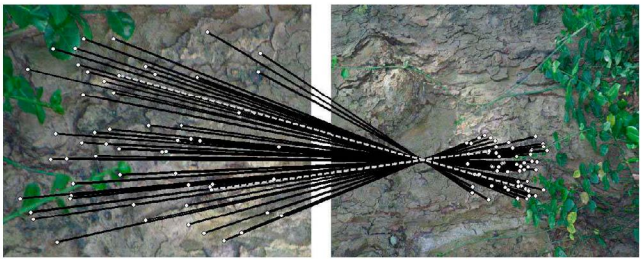
[4] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions", *IEEE Trans. PAMI*, vol. 27, no. 8, pp. 1265–1278, 2005.

[5] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *IEEE Trans. PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

(a) *Valbonne*: 17 matches, 1 incorrect (5.9%).

(b) *Valbonne*: 33 matches, 0 incorrect.

(c) *Graffiti*: 31 matches, 3 incorrect (9.7%).



(d) *Bark*: 73 matches, 2 incorrect (2.7%).

Fig. 4. Examples of retrieved correspondences (black: correct, dashed gray: false matches). (a) big occlusion and very large baseline; (b) high rotation and viewpoint change; (c) extreme viewpoint change; (d) high rotation and zoom.