A Robust Interest Points Matching Algorithm

Il-Kyun JUNG and Simon LACROIX
LAAS/CNRS
7, Ave du Colonel Roche
31077 Toulouse Cedex 4 FRANCE
Il-Kyun.Jung,Simon.Lacroix@laas.fr

Abstract

This paper presents an algorithm that matches interest points detected on a pair of grey level images taken from arbitrary points of view. First matching hypotheses are generated using a similarity measure of the interest points. Hypotheses are confirmed using local groups of interest points: group matches are based on a measure defined on an affine transformation estimate and on a correlation coefficient computed on the intensity of the interest points. Once a reliable match has been determined for a given interest point and the corresponding local group, new group matches are found by propagating the estimated affine transformation. The algorithm has been widely tested under various image transformations: it provides dense matches and is very robust to outliers, i.e. interest points generated by noise or present in only one image because of occlusions or non overlap.

1. Introduction

Many computer vision tasks rely on feature extraction and matching. For instance, feature matching can be used to compute a transformation between two images, to reconstruct the geometry of a scene (as in stereovision), or to recognize objects previously perceived. *Interest points*, *i.e.* pixels that exhibit some singularity with respect to their neighborhood, are features to which more and more attention is being paid: they can indeed be easily extracted, and are quite robust to signal noise, data acquisition parameters variations and image transformations.

One of the main application of pixel matching is stereovision. Conventional matching methods based on local grey values work well in this case, where the transformation between the two images is near the identity. Local grey value correlation techniques have been compared in [4], and an original matching technique, more robust to signal noise, has been proposed in [13]. To detect and remove false matches, a method using median flow filter has been used in [10] for the image translation case. A technique using some classical correlation and relaxation to find set of matches has been proposed in [14]. All these methods are very sensitive to changes in image scale, rotation and view point: their applications under various image transformations is therefore not possible.

To extract local characteristic stable with respect to image transformations, Gaussian derivatives and Gabor wavelet have been proposed [1]. Gabor features have been used to track facial feature points [5] : in these approaches, a high resolution of spatial frequency and angular orientation is required, which is computationally expensive. Local characteristic using Gaussian derivatives and Gauss-Laplace operator have been presented in [11]. In [7], a local characteristic vector using Gaussian derivatives is defined to retrieve images in a database.

We present here an algorithm that matches interest points between two images without any prior knowledge on the transformation between the images. It has three important properties: it is independent to various image transformations, it is not sensitive to occlusions and signal noise, and it is able to find an approximated affine transformation between the two images, in the cases where such a transformation approximates well the real transformation. It is therefore very versatile, and can be applied to image mosaicing, image retrieval in a database, object recognition, uncalibrated stereo correlation and camera motion recovery. The next section presents the interest points detection 2, and section 3 describes the heart of the algorithm, an interest point group matching procedure. Section 4 describes the strategy to establish matches over the whole image, and various results are presented in section 5.

2. Interest points

Various approaches have been proposed to detect stable interest points [6, 2, 12, 9], the most known being the Harris detector [3]. Recently, a precise version of this detector has been introduced in [8]: it uses Gaussian functions to compute the derivatives of intensity, and the two eigenvalues of the auto-correlation matrix as the principal curvatures of the auto-correlation function. The authors compared the stability of the former approaches using *repeatability*, a quantitative evaluation criteria which is the percentage of repeated interest points between two images: they assessed that the precise Harris detector provides the best repeatability.

We therefore use this version of the Harris detector: the auto-correlation matrix is

$$M = e^{-\frac{x^2+y^2}{2\sigma^2}} \otimes \begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix}$$

where I_x and I_y are computed by convolving the image with Gaussian derivatives.

The eigenvalues (λ_1, λ_2) of the matrix M are the two principle curvatures, and an interest point is declared when they are higher than a given threshold. We use a *single* scalar value as a characteristic of an interest point P, and denote it the *cornerness* c_P :

$$c_P = \| \lambda_1^2 + \lambda_2^2 \|$$

To measure the resemblance between two repeated points P and Q in two images, the *similarity* $S(P, Q)$ is defined using the cornerness c_P and c_Q:

$$S(P, Q) = \frac{min(c_P, c_Q)}{max(c_P, c_Q)}$$

Figure 2 shows the evolution of the mean of the similarity of repeated points under various known rotation and scale changes of the image shown in figure 1. Repeatability is over 80% for any rotation, and decreases down to 30% percent with a scale factor change of 1.5. For these transformations, the mean of similarity is always over 70%, and its standard deviation is not greater than 12%.

Figure 1: Interest points detected on an aerial image pair (black "+" junctions). The transformation between the images is here a 1.3 scale change, and repeatability is 63%.

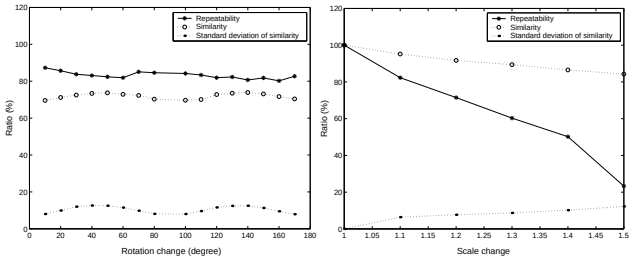

Figure 2: Evolution of the repeatability, the mean similarity and similarity standard deviation with known rotations and scale factor changes

3. Group matching procedure

We rely on three characteristics to establish matches : the first one is the *cornerness* defined on each interest point. The two other are defined on the set of interest points included in *small* local regions (referred to as *local groups*): one is the *local repeatability* of the matched groups, *i.e* the number of points that are repeated in the two groups; and the other is a correlation coefficient of the intensity of the repeated points. The local groups cover an image region small enough so that it is possible to estimate a local affine transformation between the two groups: the local repeatability is computed using this affine transformation.

Our group matching algorithm relies on the assumption that if two interest points match, their cornerness are similar and their close neighbors are also very likely to be matched together : the repeatability between matched groups is higher than for any other group pair.

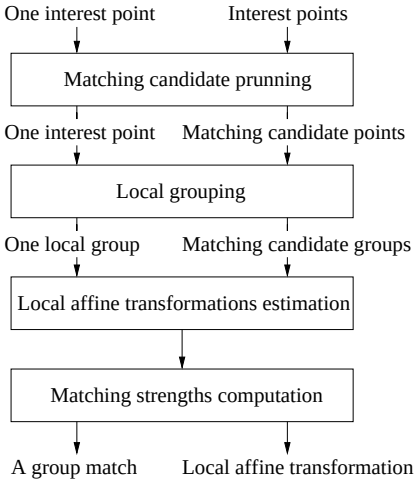

Figure 3: Group matching procedure

Figure 3 summarizes the group matching procedure : given an interest point in the first image, all the interest points in the second image whose cornerness is similar are matching candidates. To determine which candidate is the good match, local groups around the considered point and all the candidates are built. For each candidate group, the local affine transformation that yields the highest repeatability is determined. The repeatability criteria is not sufficient to establish a valid group match: group matches are confirmed by the estimation of correlation coefficient computed on the intensity of the repeated points in matched groups.

3.1. Matching candidate pruning

Let $\mathcal{P}$ the set of interest points P in the first image, and $\mathcal{Q}$ the set of interest points Q in the second image. Given a point P taken randomly in $\mathcal{P}$, the subset $\mathcal{Q}_P$ of $\mathcal{Q}$ of the possible matching candidates Q for P is defined as follows:

$$\begin{aligned} \mathcal{Q}_P &= \{Q \mid S(P, Q) < T_c\},\ Q \in \mathcal{Q} \\ &= \{Q^1, \ldots, Q^K\} \end{aligned}$$

where T_c is a threshold on cornerness similarity. The value of T_c is determined on the basis of the figures presented in section 2: supposing that the repeatability between

two matched regions of the images is higher than 50%, similarity must be greater than 70%. Empirical tests show that a value of 0.65 for T_c is satisfying.

3.2. Local grouping

Once the set $\mathcal{Q}_P$ of possible matching candidates for P is determined, the local group of interest points $\mathcal{G}_P$ around the *pivot* P is built:

$$\mathcal{G}_P = \{P, p_1 \dots p_n\}$$

$p_1 \dots p_n$ are the n closet interest points of P (the *neighbors* of P). Similarly, K local groups $\mathcal{G}_{Q^k}$ are determined for each candidate matching point Q^k in $\mathcal{Q}_P$:

$$\mathcal{G}_{Q^k} = \{Q^k, q_1^k \dots q_n^k\}$$

The neighbors are sorted according to their distance to the pivot ($\| \vec{p}_1 \| < \dots < \| \vec{p}_n \|$, where $\vec{p}_i = P - p_i$), and the number n of neighbors must be small, so that an affine transformation is a good approximation of the image transformation to match the groups. Empirically, a number of $n = 5$ neighbors gives good matching results.

3.3. Local affine transformation estimation

To compute the repeatability of interest points, one must determine the 2D transformation between the images. A method to estimate this transformation using 2D projective transformation has been proposed in [15]. But in the case of non-planar scenes, such a transformation does not exist: only a local transformation between two small regions in the images can be approximately defined. Since the interest points in a local group cover a small area, the transformation between two local groups can be approximated with scale, rotation and translation changes, the shear effects being neglected. This approximated local affine transformation between $\vec{p}_i$ and $\vec{q}_i^*$ in the groups $(\mathcal{G}_P, \mathcal{G}_{Q^k})$, is written as

$$\vec{q}_i^* = \mathbf{A}\vec{p}_i + \mathbf{t} \approx \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \vec{p}_i + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \quad (1)$$

where ρ is the scale change and θ the rotation angle. Given a group pair $(\mathcal{G}_P, \mathcal{G}_{Q^k})$, the translation $\mathbf{t}$ is first determined by subtracting the coordinates of Q^k to the coordinates of P. The matrix $\mathbf{A}$ being defined by two elements, an equivalent "affine feature vector" composed of ρ, θ is defined, and a single matching points pair is sufficient to define an affine feature vector: p_i, i_{th} neighbor of P, can be matched to q_j, j_{th} neighbor of Q if $S(p_i, q_j) < T_c$, and we have :

$$A(p_i, q_j) = (\rho(p_i, q_j), 2\hat{\theta}(p_i, q_j))$$

where

$$\rho(p_i, q_j) = \frac{\| \vec{q}_i \|}{\| \vec{p}_i \|}, \quad \hat{\theta}(p_i, q_j) = \frac{\vec{u}_{p_i} \wedge \vec{u}_{q_i}}{\vec{u}_{p_i} \cdot \vec{u}_{q_i} + 1}$$

$\vec{u}_{p_i}$ and, $\vec{u}_{q_i}$ denote respectively $\frac{\vec{p}_i}{\|\vec{p}_i\|}$ and $\frac{\vec{q}_i}{\|\vec{q}_i\|}$, and the number 1 added to the denominator of the definition of $\hat{\theta}$ avoids the undefined cases where the inner product is null.

The maximum number of possible affine vectors for a group match is n^n: one must now find out which of the possible affine vectors is the best for a tentative match between the groups $\mathcal{G}_P$ and $\mathcal{G}_{G^k}$ (denoted $\mathcal{M}(\mathcal{G}_P, \mathcal{G}_{Q^k})$). For that purpose, the possible affine vectors are grouped into classes of similar vectors. Indeed, the affine vectors between matched points in a group match should be similar, and the number of matched points in a group match should be the highest. The class of vectors similar to $A(p_i, q_j)$ is:

$$\mathcal{C}(p_i, q_j) = \{A(p_{i+u_1}, q_{j+v_1}), \dots, A(p_{i+u_N}, q_{j+v_N})\}$$

The indices in this equation are so that $u_{m+1} > u_m > 0$ and $v_{m+1} > v_m > 0$ (remember that neighbors are sorted by their distance to the pivot), and an affine vector $A(p_{i+u_m}, q_{j+v_m})$ belongs to $\mathcal{C}(p_i, q_j)$ if

$$E(u_m, v_m) = \| W(A(p_{i+u_m}, r_{j+v_m})^T - A(p_i, r_j)^T) \| < T_E$$

where W is a 2×2 orthogonal weighting matrix to adjust the contrast between variations of two affine feature elements.

After the classification, a class set $\{\mathcal{C}(p_i, r_j)\}$ is obtained. The following function is used to evaluate each class:

$$\mathcal{J}(\mathcal{C}(p_i, r_j)) = \sum_{m=1}^{N} \frac{E(u_m, v_m)}{N^2} \quad (2)$$

The affine vector class that minimizes (2) is retained for the tentative group match $\mathcal{M}(\mathcal{G}_P, \mathcal{G}_{Q^k})$,

$$\mathcal{M}(\mathcal{G}_P, \mathcal{G}_{Q^k}) = \{(P, Q^k), (p_i, q_j), \dots, (p_{i+u_m}, q_{j+v_m})\}$$

and the repeatability[1] $R_{\mathcal{M}}$ is defined for this group match, $R_{\mathcal{M}} = |\mathcal{M}(\mathcal{G}_P, \mathcal{G}_Q)|$.

3.4. Matching strength

We now have a set of local affine transformations and the corresponding repeatability for the possible group matches $\mathcal{M}(\mathcal{G}_P, \mathcal{G}_{Q^k})$, $Q^k \in \mathcal{Q}_P$. However, since we chose a small number of neighbors in a group ($n = 5$) to satisfy the assumption that an affine transformation is a good estimate of the real transformation between the groups, the repeatability $R_{\mathcal{M}}$ is not discriminative enough to distinguish good and bad group matches.

The strength of a group match is therefore defined as a combination of repeatability $R_{\mathcal{M}}$ and a correlation score of intensity between repeated points in the groups $\mathcal{G}_P$ and $\mathcal{G}_{Q^k}$. We chose the zero-mean normalized cross-correlation score (ZNCC):

[1] repeatability is here an the integer number of repeated points, not a percentage ratio.

$$Z(\mathcal{M}(\mathcal{G}_P, \mathcal{G}_{Q^k})) = \frac{\sum_{i=1}^{R_\mathcal{M}} (I_{p_i} - \overline{I}_p)(I_{q_i} - \overline{I}_q)}{\left[\sum_{i=1}^{R_\mathcal{M}} (I_{p_i} - \overline{I}_p)^2 \sum_{i=1}^{R_\mathcal{M}} (I_{q_i} - \overline{I}_q)^2\right]^{1/2}}$$

To cope for the noise in the images, the ZNCC score is computed on the average intensity of a small (3×3) local windows around the repeated interest points. The strength of a match $\mathcal{M}$ is defined as follows:

$$\mathcal{S}(\mathcal{M}) = \begin{cases} R_\mathcal{M} + Z(\mathcal{M}) & \text{if } \mathrm{Z}(\mathcal{M}) > T_Z \\ 0 & \text{else} \end{cases}$$

where T_Z is a threshold on $Z(\mathcal{M})$ and T_Z have to be positive. $R_\mathcal{M}$ is the number of repeated points and the maximum value of $Z(\mathcal{M})$ is 1: with this definition, the group match which has the highest repeatability and a ZNCC score higher than T_Z is considered valid, and if two group matches have the same repeatability $R_\mathcal{M}$, then they are qualified by the ZNCC score.

4. Interest points matching algorithm

Once a reliable group match is found by group matching procedure, a focussed group matching procedure is activated. Although the local affine transformation found for a group match is not globally stable on most scenes, it is locally stable. Therefore, it is propagated around the current group match found , in order to focus the search of match candidates, thus reducing both the number of candidates to check and the possibility of false matches occurrences. The whole matching algorithm procedure is depicted in figure 4.

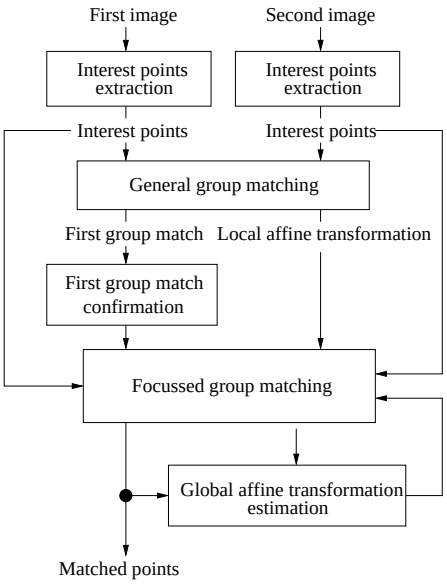

Figure 4: Interest points matching algorithm

4.1. First matching group confirmation

Since further matches will be established on the basis of the local affine transformation of the current match, the validity of the first group match is of extreme importance. To make sure that the first group match is a valid one, matches are established for a few groups near the first matched group: if the found match is a real one, matches for very close groups must have a very similar local affine transformation. The similarity between the local affine transformations defined by these new group matches and by the first one is evaluated. A local affine vector $\mathcal{A} = [\, a_{11},\ a_{12},\ a_{21},\ a_{22},\ t_x,\ t_y \,]$ that represents the affine transformation of equation 1 is defined, and the similarity between two affine vectors is qualified by the Mahalanobis distance:

$$d_M(\mathcal{A}_i, \mathcal{A}_j) = \sqrt{(\mathcal{A}_i - \mathcal{A}_j)^T \wedge^{-1} (\mathcal{A}_i - \mathcal{A}_j)} \quad (3)$$

If more than one neighbor group match produces a local affine vector whose Mahalanobis with the first group vector is below a threshold, the first group match is declared as a valid one. If not, an other interest point is randomly selected in the first image, and the process is reiterated until a valid group match is found.

4.2. Focussed group matching

The local affine transformation obtained by the first group match is only valid near this group: it is used to estimate a search window in the second image to match the interest points close to the first group in the first image. Given a matched group $\mathcal{G}_P$ in the first image, the next interest points P' to study are:

$$\{P'\} = \{P' \mid (d(P', P) < T_d) \wedge (\mathcal{A}(P') \in Im_2)\} \quad (4)$$

where $d(P', P)$ is a distance between two interest points in the first image, T_d is a threshold on this distance, and $\mathcal{A}(P')$ are the coordinates of P' in the second image after the application of the current local affine transformation. The matching candidates for a point P' are

$$\mathcal{Q}_{P'} = \{Q \in Im_2 \mid Q \in \mathcal{W}_{\mathcal{A}(P')}\} \quad (5)$$

where $\mathcal{W}_{\mathcal{A}(P')}$ is a search window around $\mathcal{A}(P')$. P' and the (small) set of matching candidates are then fed to the group matching procedure described in section 3. The local affine transformation found after the production of a new group match is then checked: its similarity to the transformation of the former match is tested with the Mahalanobis distance (3). If the new match is confirmed by this test, the current local affine transformation is updated and the process is repeated until all interest points in the first image are studied.

4.3. Global affine transformation

While the local affine transformation estimate is propagated and updated, it is possible to estimate a global affine

transformation, which is of course an approximation of the real transformation between the images. But the global transformation gives much precise coordinate estimation $\mathcal{A}(P')$ in case of almost planar scenes or small view point change: it can be used to study the unmatched points after the end of the focussed search. A least square technique is used to compute the global affine transformation using (1), by minimizing the accumulated error e:

$$e = \sum_i (\mathbf{P}_i \mathbf{x} - \mathbf{Q}_i)^2$$

where $\mathbf{P}_i$, $\mathbf{Q}_i$ are the i_{th} matching points and $\mathbf{x}$ is the affine model parameters. When the global affine transformation become stable, it is used as the transformation $\mathcal{A}$ in equations (4) and (5).

5. Results

We tested our algorithm with a lot of distortion free image pairs of various 3D scenes, taken from positions that yields various complex image transformations. To define the effectiveness of the algorithm, two kinds of evaluation method are used: for planar scenes, a homography $\mathcal{T}$ is determined thanks to a non-linear constraint least square method applied on the matched interest points coordinates. The distance between a point p and its match q is defined by $\| q - \tilde{p} \|$, where $\tilde{p} = \mathcal{T}(p)$. For non-planar scenes, results are evaluated by the distance between the epipolar line of a point and its matching point. The fundamental matrix F that defines the epipolar lines is estimated by a least median square method [14]. The distance is then defined as follows:

$$d(p,q) = \frac{1}{2}(\| p - \tilde{p} \| + \| q - \tilde{q} \|) \qquad (6)$$

where $\tilde{p}, \tilde{q}$ are the points on the corresponding epipolar lines: $\tilde{q} F p = 0$, $\tilde{p} F^T q = 0$ and the lines defined by $(p, \tilde{p}), (q, \tilde{q})$ are orthogonal to the epipolar lines.

In the following figures, "wrong matches" are represented by a cross, "good matches" by a square. For planar scenes, the tolerance limit of the distances is set as the summation of the mean distance and two times the standard deviation of their distribution[2], and for non-planar scenes the tolerance is set to 1.5 pixel.

Figure 5 presents the matching results for an aerial image pair. In this case, the elevation of the camera is much bigger than the scene depth variation: the scene can be considered planar, and the projective transformation defines a precise relation: the mean and standard deviation of the distances are very small (1.35 and 1.07 pixel). In this example, computation time is $2.3s$[3].

The scene of the images of figure 6 is obviously not planar: false matches are detected using the fundamental matrix (computation time is here $0.7s$). The aerial images of figure 7 have been taken from an altitude of about $20m$ by a camera mounted on a balloon. Although the image quality is poor (due to a long video cable), only 7 matches are wrong out of 68 good matches. The images in figure 8 are down-sampled to a 192×256 size, which does not disturb at all the algorithm, since it only reduces the number of interest points.

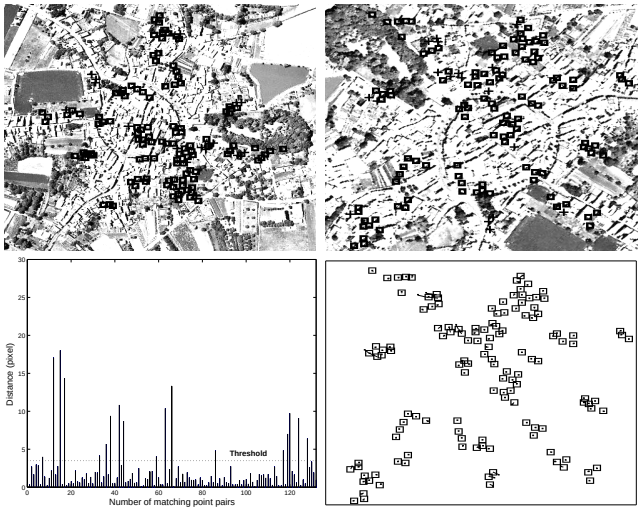

Figure 5: Matching results for a 500×500 aerial image pair, with a 140^o rotation and 1.4 scale change. The bottom right image represents the matched interest points of the top right image (squares) with the top left image matched points (dots), after the application of the projective transformation. The distribution of the distances between matched points after the application of this transformation is displayed on the bottom left. 86.4% of the matches are valid here.

Finally, figure 9 shows the results on a 480×640 indoor scene image pair. No matches have been established on the man who moved between the acquisitions: indeed, no matches consistent with the affine transformation estimates were possible do establish. Finally, we tested various cases with no overlap between the two images: no matches were ever produced.

The time performance of the algorithm depends on the size of the overlap region between the two images: when the overlap is big (as in stereo pairs for instance), a first match is quickly established. But when the overlap is very small, the algorithm can spend quite a lot of time to establish the first group match, by checking randomly interest points until a good match is found. Once a first group match is established, the remaining time is proportional to the number of actual matches in the image. Of course, time performance is dramatically enhanced if an initial estimate of the image transformation is available.

6. Summary

We presented an interest point matching algorithm which is robust under various image transformations. It relies on a single characteristic defined on interest points and on local groups of interest points. Cornerness similarity and a measure based on repeatability and intensity correlation

[2] Note that if a scene is not completely planar, a more flexible tolerance limit can be applied according to the distribution

[3] Computation times are estimated on a Sun Ultra 10 Sparc station throughout this section

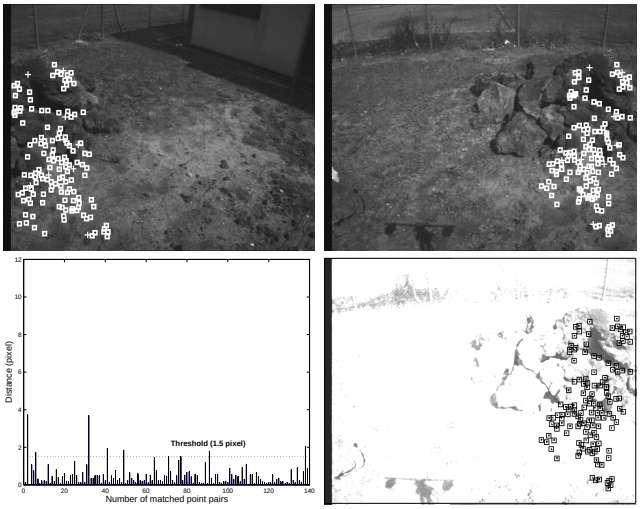

Figure 6: Matching of a 576×768 image pair with a small translation and a big panoramic rotation. 93.6% of the matches are good matches.

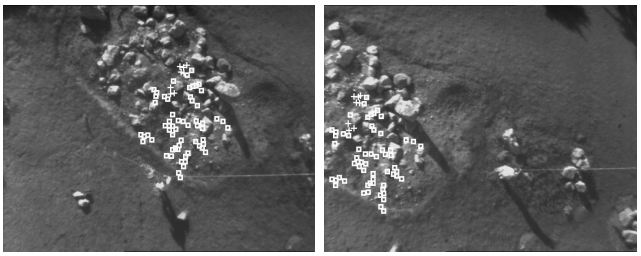

Figure 7: Matching of 576×768 aerial image pair, with an important rotation and small scale, translation and view point changes. 89.7% of the matches are good matches, the computation time is $4.3s$.

score between repeated interest points is used to match local groups. To define repeatability among local groups, a local affine transformation estimation method is proposed, and to assess the validity of a group match, a matching strength that combines the two groups characteristics is defined. Once the algorithm is initialized, an efficient candidate selection based on the local affine transformation focuses the match search. Tests on real images have shown good matching results: wrong matches seldom exceed 20%, and the algorithm works in very different conditions.

References

[1] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, A(2):1160–1169, 1985.

[2] W. Förstner. A framework for low level feature extraction. In *European Confenrence on Computer Vision*, 1998.

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

[4] J. Martin and J. L. Crowley. Comparison of correlation techniques. In *Intelligent Autonomous Systems*, 1995.

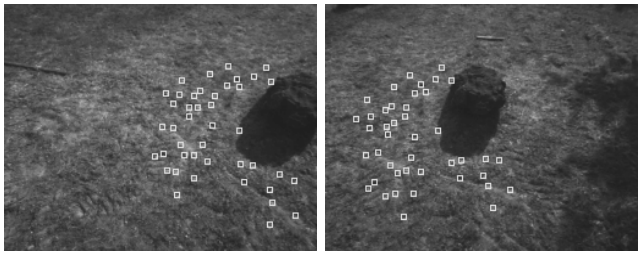

Figure 8: Matching a 192×256 image pair, 100% of the matches are valid. Computation time is here only $0.07s$.

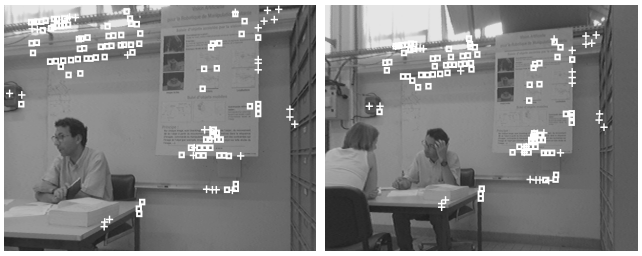

Figure 9: Matching with scale and view point change, and moving elements in the scene. 73.9% of the matches are valid, computation time is $0.6s$.

[5] D. Nanin and Al. Tracking facial feature points with gabor wavelets and shape models. In *International Conference on Audio-and Video-based biometric Person Authentication*, pages 35–42, 1997.

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

[7] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1999.

[8] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. *Proceeding of the 6th International Conference on Computer Vision*, pages 230–235, 1998.

[9] E. Shilat, M. Werman, and Y. Gdalyahu. Ridge's corner detection and correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 976–981, 1997.

[10] P. Smith and Al. Effective corner matching. In *British Machine Vision Conference*, 1998.

[11] T. Starner, B. Schiele, and A. Pentland. Visual contestual awareness in wearable computing. Technical Report 452, MIT Media Laboratory, 1998.

[12] H. Wang and M. Brady. Real-time coner detection algorithm for motion estimation. *Image and Vision Computing*, 13(9):695–703, 1995.

[13] R. Zabih and J. Woodfill. A non-parametric approach to visual correspondence. In *IEEE PAMI*, 1998.

[14] Z. Zhang and Al. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical Report 2273, INRIA, May 1994.

[15] I. Zoghalmi, O. Faugeras, and R. Deriche. Using geometric corners to build a 2d mosaic from a set of images. In *IEEE Conference on Computer Vision and Pattern Recognition, San Juan (Porto Rico)*, pages 420–425. INRIA Sophia, June 1999.