Fast and Robust Image Matching using Contextual Information and Relaxation

Dro Desire Sidibe, Philippe Montesinos, Stefan Janaqi

FAST AND ROBUST IMAGE MATCHING USING CONTEXTUAL INFORMATION AND RELAXATION

Desire Sidibe, Philippe Montesinos, Stefan Janaqi
LGI2P/EMA - Ales School of Mines, Parc scientifique G. Besse, 30035 Nimes Cedex 1, France
{*Desire.Sidibe, Philippe.Montesinos, Stefan.Janaqi*}*@ema.fr*

Keywords: Relaxation, Image matching, Point matching, Scale invariant features.

Abstract: This paper tackles the difficult, but fundamental, problem of image matching under projective transformation. Recently, several algorithms capable of handling large changes of viewpoint as well as large scale changes have been proposed. They are based on the comparison of local, invariants descriptors which are robust to these transformations. However, since no image descriptor is robust enough to avoid mismatches, an additional step of outliers rejection is often needed. The accuracy of which strongly depends on the number of mismatches. In this paper, we show that the matching process can be made robust to ensure a very few number of mismatches based on a relaxation labeling technique. The main contribution of this work is in providing an efficient and fast implementation of a relaxation method which can deal with large sets of features. Futhermore, we show how the contextual information can be obtained and used in this robust and fast algorithm. Experiments with real data and comparison with other matching methods, clearly show the improvements in the matching results.

1 INTRODUCTION

The problem of finding correspondences between image features is fundamental in many computer vision applications such as stereo-vision, image retrieval, image registration, robot localization and object recognition. Recently, local and invariant features have proven to be very successful in establishing image-to-image correspondences. The local character yields robutsness to occlusion and varying background, and invariance makes them robust to scale and viewpoint changes. Interest points are one of the most widely used local features. In many applications, one aims to obtain a set of corresponding points between two images. Therefore, the extracted points have to be characterized by a descriptor and then matched using a similarity measure.

Different methods for detecting invariant features are proposed (Baumberg, 2000; Mikolajczyk and Schmid, 2002; Tuytelaars and Van Gool, 2004; Lowe, 1999; Schaffalitzky and Zisserman, 2002; Matas et al., 2002). Among them, it is worth mentioning those based on interest points. (Mikolajczyk and Schmid, 2002; Mikolajczyk and Schmid, 2004) propose a scale and affine invariant interest points detector using a scale-space representation of the image. First, points are detected at multiple scales using the Harris detector. Then points at which a local measure of variation is maximal over scales are selected. Finally, an iterative algorithm modifies location, scale and local shape of each point and converges to affine invariant points. Scale-space representation is also used by (Lowe, 1999) who uses local extrema of Difference-of-Gaussian (DoG) filters as key-points. Similar ideas are used by other authors (Baumberg, 2000; Schaffalitzky and Zisserman, 2002). For a more detailed review on affine invariant features detection, please refer to (Mikolajczyk et al., 2005).

Once the points are detected, the region around each of them is used to compute a descriptor. Invariance to affine transformations is provided by the fact that each point is characterized by a specific scale which defines the size of its region and that each region has a specific shape. Many different techniques for describing local image regions have been developed and it has been shown that the SIFT (Scale and Invariant Feature Transform) descriptor performs bet-

ter than others (Mikolajczyk and Schmid, 2005). This descriptor is based on the gradient distribution in the detected regions around the points and is represented by a 3D histogram of gradient locations and orientations (Lowe, 1999).

Affine invariant points combined with a distinctive descriptor such as SIFT lead to very good results in the presence of significant transformations. However, while in the aforementioned works much effort is done for computing distinctive descriptors, less attention is paid to the matching strategy. A simple comparison of the descriptors, for example using Euclidean or Mahalanobis distance, and matching to nearest neighbour will always give some mismatches. This is because no image descriptor is robust enough to be perfectly discriminant and avoid mismatches. Thus, an additional step of outliers rejection is often needed. One approach is to estimate the geometric transformation between the pair of images and use this information to reject inconsistent matches (Zhang et al., 1995). This can, of course, be done only in stereo-vision or in matching images containing planar structures for which the epipolar constraint or a plane homography can be estimated. The accuracy of the estimation relies on the number of mismatches. This number can be reduced by considering the ratio between the first and second nearest neighbour, i.e. matching a point to its nearest neighbour if this one is much more closer than the second nearest neighbour (Zhang et al., 1995; Lowe, 1999). Taking into acount a kind of ambiguity measure, this strategy reduces the number of mismatches. Unfortunately, it reduces the number of correct matches as well.

Moreover, when the ambiguity is high as it is in the presence of repetitive patterns, see Figure 1, the previous methods fail to find correct matches. That is because, in these cases, all the points have almost the same SIFT descriptor. So, matching to nearest neighbour gives a lot of mismatches. Taking some additional information into account during the matching process could reduce the ambiguity. This is the main idea of the widely used relaxation labeling technique. However, most of the existing algorithms (Rosenfeld et al., 1976; Faugeras and Berthod, 1981) have prohibitive complexity and are therefore limited to the assignment of a small number of labels.

In this paper, we present a matching method based on relaxation which can handle large point sets and provide a very few number of mismatches under important transformations. This work is based on an algorithm presented by (Faugeras and Berthod, 1981) and our main contribution is in providing a fast and efficient implementation of this algorithm. Futhermore, we show how the contextual information can be obtained and used in this robust and fast algorithm. The remainder of the paper is organized as follows. In Section 2, we decribe the relaxation labeling techniques and show their limits. Then our efficient implementaion is given in Section 3. Experimental results showing the improvements of the method over other existing techniques are presented in Section 4. Finally, concluding remarks are given in Section 5.

Figure 1: A difficult case of matching. Matching to nearest neighbour fails because of repetitive patterns.

2 RELAXATION MATCHING

2.1 Relaxation Labeling Techniques

The relaxation labeling technique was first introduced by (Rosenfeld et al., 1976) to deal with ambiguity and noise in vision system. Let $u = \{u_1, \ldots, u_n\}$ and $v = \{v_1, \ldots, v_m\}$ be two sets of points from two images. Each point is characterized by a descriptor. The principal idea of relaxation is to use the information provided by the neighbourhood of each point to improve consitency and reduce ambiguity. More precisely, let define for each point u_i a set of initial probabilities $p_i^0(k), k = 1, ..., m$; $p_i^0(k)$ being the probability that point u_i is matched with point v_k. An iterative process is designed to update the probabilities until a consistent distribution is reached. The update is based on a support, or compatibility, function q_i defined in the neighbourhood V_i of the point u_i. This support function measures the likelihood of a point u_i to be matched with a point v_k, given the configuration of its neighbours. Many probabilistic relaxation schemes have been proposed and they essentially differ in the definition of the support function and the updating rule. For example, one standard updating rule is defined by (Hummel and Zucker, 1983) as:

$$p_i^{t+1}(k) = \frac{p_i^t(k) q_i^t(k)}{\sum_k p_i^t(k) q_i^t(k)} \qquad (1)$$

where

$$q_i^t(k) = \sum_j w_{ij} \left[\sum_l p_{ij}(k,l) p_j^t(l) \right] \quad (2)$$

and $p_{ij}(k,l)$ is the probability that point u_i is matched with point v_k under the condition that point u_j is matched with v_l. $p_{ij}(k,l)$ is the contextual information that helps improving consistency. The scalars w_{ij} are weights that indicate the influence of point u_j on point u_i. They are normalized and verify $\sum_j w_{ij} = 1$.

(Faugeras and Berthod, 1981) propose a relaxation scheme based on an optimization approach. They define a global criterion to be minimized considering both consistency and ambiguity:

$$C = \alpha C_1 + (1-\alpha) C_2 \quad (3)$$

where the consistency measure is:

$$C_1 = \frac{1}{2n} \sum_{i=1}^{n} \|p_i - q_i\|^2 \quad (4)$$

and the ambiguity measure is:

$$C_2 = \frac{m}{m-1} \left[1 - \frac{1}{n} \sum_{i=1}^{n} \|p_i\|^2 \right] \quad (5)$$

Let x be the vector obtained by concatenating the vectors p_i, i.e. $x = [p_1, \ldots, p_n]^T$. Then, the problem of finding a set of corresponding points comes down to minimizing $C(x)$ subject to the linear constraints:

$$\begin{cases} \sum_{k=1}^{n} x_i(k) = 1 \quad i = 1, \ldots, n \\ x_i(k) \geq 0 \quad i = 1, \ldots, n \quad k = 1, \ldots, m \end{cases} \quad (6)$$

The optimization problem is solved by a projected gradient method and for each point u_i, the point v_k with highest final probability is retained as its correspondent. This approach seems better since the final set of matches will be more consistent and less ambiguous. However, it is limited in practice by its high complexity.

2.2 Drawbacks of the Original Method

The main limitation of the optimization approach (Faugeras and Berthod, 1981) and the nonlinear approaches (Rosenfeld et al., 1976; Hummel and Zucker, 1983) is their high complexity. The former is in fact a $O(nm^2V)$ algorithm where V is the size of V_i for $i = 1, \ldots, n$. Thus, these algortihms are appropriate to applications such as image segmentation or classification issues where one needs to assign a small number of labels, i.e. $m \approx O(10^2)$. For applications such as image matching where one needs to assign a large number of points from one image to the other, $m \approx O(10^4)$, the methods become impractical. This is mainly because the compatibility function q_i (see Equation 2) has to be re-estimated at each iteration.

Another limitation is the fact that the final probabilities critically depend on the initial and the conditional probabilities (Hummel and Zucker, 1983; Price, 1985). If these quantities are not correctly estimated, then the final probabilities will provide a lot of mismatches.

In the next section we address these two problems and we show how the complexity can be considerably reduced in order to handle large point sets. We also provide a way of computing the conditional probabilities.

3 FAST AND EFFICIENT MATCHING ALGORITHM

3.1 Reducing the Complexity

In order to reduce the complexity of the optimization approach, we show that the criterion C of Equation 3 can be written in the following form:

$$C(x) = \frac{1}{2} x^T H x + cte \quad (7)$$

i.e.

$$C([x_1, \ldots, x_n]^T) = \frac{1}{2} \sum_{t=1}^{n} \sum_{p=1}^{n} x_t^T H_{tp} x_p + cte \quad (8)$$

where

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & H_{ij} & \vdots \\ H_{n1} & \cdots & H_{nn} \end{pmatrix}$$

and each matrix H_{ij} contains the conditional probabilities $p_{ij}(k,l)$, i.e. the contextual information needed to compute the support function q_i. See appendix for details about obtaining the matrices H_{ij}.

Firstly, if we consider in the definition of the support function (Equation 2) only points u_j which are in the neighbourhood V_i of point u_i, then it is clear that some of the matrices H_{ij} are equal to zero. In particular, it is easy to show that for $i = 1, \ldots, n$ and for $j = 1, \ldots, n$:

$$H_{ij} \neq 0 \quad \text{if} \quad \begin{cases} i = j \quad \text{or} \\ u_j \in V_i \quad \text{or} \\ \exists k / (u_i, u_j) \in V_k \times V_k \end{cases} \quad (9)$$

Therefore, using a sparse matrix representation for H we reduce the complexity of the method. To reduce

the complexity further, we bring down the set of potential matches for each point u_i to the set of its K nearest neighbours given a similarity measure. Thus, each matrix H_{ij} is of size $K \times K$ instead of $m \times m$. With $K \ll m$, this reduces memory requirement of the algorithm.

Secondly, the matrix H is computed only once and that makes the algorithm faster. At each iteration the gradient of the criterion is obtained by the following equation:

$$\frac{\partial C}{\partial x} = \frac{1}{2}(H + H^T)x \tag{10}$$

In general, H is not a symmetric matrix. But in the case it is, the gradient is given by the classical equation:

$$\frac{\partial C}{\partial x} = Hx \tag{11}$$

3.2 Initial and Conditional Probabilities

We mentioned already, see Section 2.2, that the results, i.e. the final probabilities, of a relaxation scheme critically depend on the initial probabilities and the conditional probabilities. So, estimation of these quantities is of great interest.

Initial probabilities are computed based on Euclidean distance between descriptors. We used SIFT as it is considered to be the best local descriptor (Mikolajczyk and Schmid, 2005) and we choose the K nearest neighbours points v_k as the potential matches for each point u_i. Then, the initial probabilities are given by the following equation:

$$p_i^0(k) = \frac{1/d_{ik}}{\sum_{k=1}^{K} 1/d_{ik}} \quad i = 1,\ldots,n \quad k = 1,\ldots,K \tag{12}$$

where d_{ik} is the Euclidean distance between the descriptors of points u_i and v_k.

For each interest point u_i, the compatibility function q_i indicates how a match assigned to point u_i is *consistent* with those of its neighbours in V_i. Thus, q_i can be seen as an estimation of p_i given the prior knowledge represented by the $p_{ij}(k,l)$ for points u_j in V_i. Estimation of the $p_{ij}(k,l)$ can be done using geometric and photometric information of the scene. Geometric semi-local constraints are used by (Schmid and Mohr, 1997; Montesinos et al., 2000; Tuytelaars and Van Gool, 2004). In (Pelillo and Refice, 1994) the compatibility coefficients are learned from training examples. We based the estimation of our contextual information on photometric information because in case of large viewpoint changes, geometry is badly preserved. Moreover, as the SIFT descriptor gives a geometric description of a point's neirghbourhood, it makes sense to use a complementary photometric information for matching.

For each point u_i and each of its neighbours $u_j \in V_i$, we define a rectangular patch M_{ij} of length l_{ij} and width $l_{ij}/2$, l_{ij} being the distance between u_i and u_j. Note that in order to discard very small patches, we consider in V_i only points u_j which are at a distance from u_i greater than $5\sigma_i$, σ_i being the specific scale of point u_i:

$$u_j \in V_i \quad \text{if} \quad d_{ij} \geq 5\sigma_i \tag{13}$$

This is because the detector can find two or more points that are at the same location but with different orientations. Each patch is normalized to a unit square, and conditional probabilities are computed as normalized cross-correlation between patches in both images.

3.3 Matching Strategy

In many applications, one-to-one correspondence is desired. But in general, because of occlusions, varying background and scale and viewpoint changes, not all points in u will have a correspondance in v. To solve this problem, one adds a *nil* point, v_{nil}, to the set of potential matches of each point u_i. Thus, for each point $u_i \in u$, the set of potential matches is:

$$PM_i = \{v_1^i, \ldots, v_K^i, v_{nil}\} \tag{14}$$

where $v_1^i, \ldots, v_K^i$ are the K nearest neighbours points v_k of u_i based on the Euclidean distance between SIFT descriptors as described in Section 3.2. The matrices H_{ij} are of size $(K+1) \times (K+1)$ and can be written as follows:

$$H_{ij} = \left(\begin{array}{ccc|c} p_{ij}(1,1) & \cdots & p_{ij}(1,K) & p^{**} \\ \vdots & & \vdots & \vdots \\ p_{ij}(K,1) & \cdots & p_{ij}(K,K) & p^{**} \\ \hline p^{**} & \cdots & p^{**} & p^{**} \end{array}\right) \tag{15}$$

where p^{**} is a constant value defining the conditional probabilities for v_{nil}. Initial probabilities for v_{nil} are also set to a constant value:

$$p_i^0(v_{nil}) = p^* \quad i = 1,\ldots,n \tag{16}$$

Once the matrices H_{ij} are obtained, the matrix H is computed (see Section 3.1) and the optimization problem is solved by a projected gradient method. The algorithm converges to a local minimum after a reduced number of iterations and for each point u_i, the potential match with highest final probability is retained as its correspondent. Points in one image which have no correspondent in the other image are expected to match with v_{nil}.

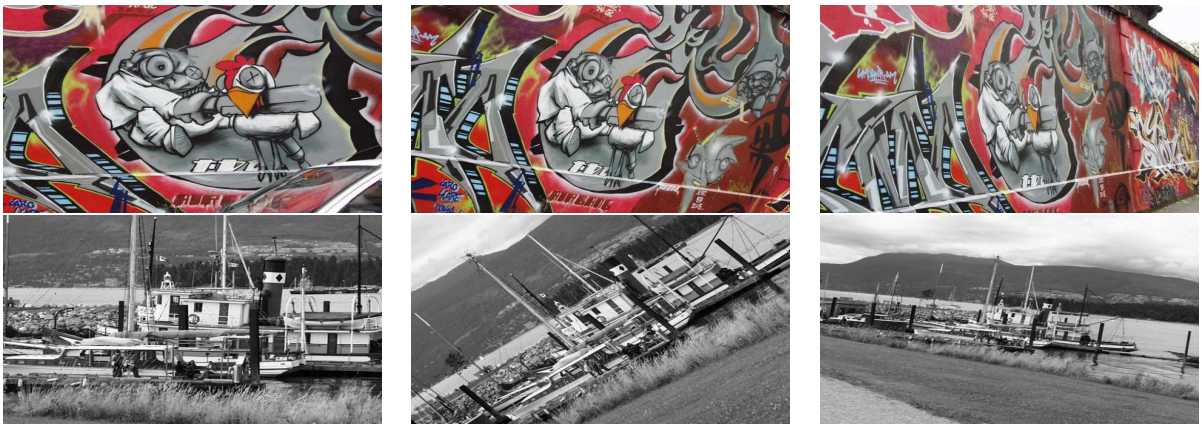

Figure 2: Examples of images used for wide baseline matching. Top: first, third and fifth frame of the *Graf* sequence. Bottom: first, third and fifth frame of the *Boat* sequence.

4 EXPERIMENTAL RESULTS

In this section we report some experiments carried out on real images to evaluate the performance of the algorithm. First, we have conducted experiments in the case of matching with large baseline using some images from (Mikolajczyk and Schmid, 2005) [1] and the images presented in Figure 1. Secondly, we apply the method to features-based object recognition. In all our experiments we set $V = K = 5$, i.e. each point has 5 neighbours and 5 potential matches. We set the constant α of the criterion C (Equation 3) to 0.5, i.e. consistency and ambiguity measures are given the same importance. And initial and conditional probabilities for the *nil* point are taken equal to 0.1, i.e. $p^* = p^{**} = 0.1$.

We compare our method, named ORELAX for optimization approach, with the three following techniques:

- CRELAX: relaxation technique using the classical updating rule defined in Equation 1 (Hummel and Zucker, 1983);
- NNDR: nearest neighbour distance ratio (Lowe, 1999). That is a point is matched to its nearest neighbour if this one is much more closer than the second nearest neighbour:

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

  where $D_i = \{d_{il}, l = 1, \ldots, m\}$;
- SVD: a SVD-based method using SIFT features (Delponte et al., 2006). A proximity matrix G is computed using SIFT descriptors of points, and matches are found based on a SVD decomposition of G:

  $$G_{ij} = e^{-d_{ij}^2/2\sigma^2}$$

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

4.1 Wide Baseline Matching

For these experiments the homographies between differents views are available and we can compute the matching rate (MR) of a matching method as the ratio between the number of correct matches and the number of detected matches.

$$MR = \frac{\#\ of\ correct\ matches}{\#\ of\ detected\ matches} \tag{17}$$

A couple of corresponding points (p, p') is said to be a correct match if:

$$\|p' - \mathcal{H}p\| < 5 \tag{18}$$

where $\mathcal{H}$ is the homography between the two images.

We use two sequences, the *Graf* and the *Boat* sequences for evaluation. The former is a six frames sequence with important viewpoint change between the first and the following frames, and the later is a six frames sequence with rotation and scale change. Some of these images are shown in Figure 2.

We present an example of matching results obtained by our method for the *Graf* sequence in Figure 3. There are respectively 1401 and 1279 interest points detected in the first and fourth frames. The algorithm finds 82 matches between these two frames and 71 of them are correct.

Figure 3: Example of matching results with the *Graf* sequence: 71 corrects matches are found between the first and fourth frames.

Table 1 shows comparative results obtained with the different methods. It can be seen that for this diffi-

cult case, ORELAX gives considerably more matches than NNDR while maintining a high MR (matching rate). SVD and CRELAX also provide a large number of matches but with a very poor MR, less than 0.5. This is mainly because the SVD decomposition algorithm has stability problems when dealing with large matrices. We used the algorithm implemented in MATLAB for our experiments. Therefore, instead of improving results, SVD based method spoils the results obtained by simple descriptors comparison as with NNDR. Moreover, ORELAX is much faster than SVD. It is slower than NNDR since the later simply compares Euclidean distance between descriptors. ORELAX is faster than CRELAX because of the optimization step which lead to a fewer number of iterations for the former.

Table 1: Comparison of different algorithms using the first and fourth frames of the *Graf* sequence.

Methods	# of matches	# of correct	MR	time in *s*
ORELAX	82	71	0.86	8.56
NNDR	35	26	0.74	2.9
SVD	118	58	0.49	47.1
CRELAX	84	34	0.40	11.41

Additional results for the whole *Graf* sequence are shown in Table 3. We see that the relaxation method with optimization gives better results for varying viewpoints. ORELAX returns the highest number of matches with the highest MR. NNDR gives the second best performance but it returns about twice less matches than ORELAX. SVD and CRELAX have poor MR for the last three frames. The number of matches goes down sensibly when viewpoint change becomes important. For example, viewpoint change between the first and fifth frames of the *Graf* sequence is greater than 50 degres. The SIFT descriptor cannot cope with such large viewpoint as reported in (Mikolajczyk and Schmid, 2005; Delponte et al., 2006).

Results obtained for the *Boat* sequence are presented in Table 4. They are similar to those obtained with the *Graf* sequence but there are more correct matches for the last frames. Moreover, the MR of ORELAX and NNDR is almost always equal to 1 except for the last frame. This means that the SIFT descriptor is more suited to rotation and scale changes than to large viewpoint changes.

In the case of repetitive patterns as in Figure 1, the relaxation method with optimization gives better results. Results presented in Table 2 show that a simple comparison of descriptor fails to find enough correct matches. The number of matches provided by NNDR is too small and the the proportion of outliers obtained by SVD is too high. Therefore, an estimation of the geometric transformation by a method such as RANSAC will fail. On the contrary, ORELAX gives almost six times more matches than NNDR with a high MR. It is important to emphasize that in this difficult case, one should consider a high number of potential matches for each point to reduce ambiguity. We set $K = 7$. See Section 4.3 for a discussion about the influence of algorithm's parameters.

Table 2: Comparison of different algorithms using the images in Figure 1.

Methods	# of matches	# of correct matches	MR
ORELAX	38	25	0.66
NNDR	6	3	0.50
SVD	60	21	0.35
CRELAX	120	26	0.22

4.2 Object Recognition

We also compare the different algorithms in a case of features-based object recognition. The results are shown in Figure 4. For the first experiment, a book is placed on a desktop such that it is partially occluded and has its scale, orientation and viewpoint changed. There are respectively 193 and 2541 interest points detected on the object, shown in the top of Figure 4a, and on the entire scene shown in the bottom of Figure 4a. ORELAX finds 20 matches, all of which are corrects, while NNDR finds only 7. CRELAX finds 8 correct matches and SVD gives 12 correct matches over 18 detected matches.

For the more difficult case shown in Figure 4b, ORELAX finds 5 correct matches over 8 detected matches, while the other methods fail to find any correct matches. Note that the images are shown at their actual relative scale.

4.3 Influence of algorithm's parameters

Results depend on the values of the algorithm's parameters. It is clear that the greater V and K are, the more accurate the method will be at the cost of a longer processing time.

To measure the influence of the parameter α, we use the same pair of images. The results presented in Table 5 show that if more importance is given to the consistency term of the criterion, i.e. $\alpha > 0.5$, then the MR of the method increases but the number of matches decreases. On the contrary, if more importance is given to the ambiguity term, i.e. $\alpha < 0.5$,

Table 3: Comparison of different algorithms using the *Graf* sequence.

Frame number	ORELAX		NNDR		SVD		CRELAX	
	# of matches	MR	# of matches	MR	# of matches	MR	# of matches	MR
2	530	0.98	261	0.97	363	0.89	384	0.94
3	180	0.93	91	0.70	222	0.67	198	0.58
4	82	0.86	35	0.74	118	0.49	84	0.40
5	11	0.72	3	0.67	60	0.13	35	0.03
6	5	0.6	9	0	40	0.1	27	0

Table 4: Comparison of different algorithms using the *Boat* sequence.

Frame number	ORELAX		NNDR		SVD		CRELAX	
	# of matches	MR	# of matches	MR	# of matches	MR	# of matches	MR
2	620	0.99	294	0.98	427	0.91	490	0.93
3	488	0.99	183	0.99	365	0.87	289	0.94
4	127	0.99	58	0.99	125	0.84	97	0.45
5	75	0.99	46	0.99	76	0.83	79	0.3
6	8	0.75	6	0.5	47	0.45	46	0.08

a)

b)

Figure 4: Examples of recognition results. a): ORELAX finds 20 matches all of which are correct. b): ORELAX finds 8 matches and 5 of them are correct.

then the number of matches increases while the MR of the method decreases.

There is a balance to find between the matching rate and the number of detected matches. We have found that the values $V = k = 5$ ($K = 7$ in the case of repetitive patterns) were sufficient for our experiments. The results presented in the previous sections are obtained with $\alpha = 0.5$, i.e. consistency and ambiguity measures are given the same importance.

Table 5: Influence of the parameter α using the images of Figure 1.

α	# of matches	# of correct matches	MR
0.3	88	42	0.48
0.5	38	25	0.66
0.7	23	17	0.74
0.9	15	13	0.87

5 CONCLUSION

In this paper a fast and robust image matching method is proposed. The method is based on relaxation labeling technique and optimization. We showed that writing the criterion to mimize in a convenient way and using a distinctive descriptor such as SIFT, the complexity of the algorithm can be considerably reduced. Futhermore, we showed how the necessary contextual information can be obtained in order to improve matching results and reduce the number of mismatches. Experimental results in case of wide baseline and in case of object recognition show that this approach gives superior results compared with other matching methods. Roughly speaking, we gain at least 30% on the number of matches and the number of correct matches as well. We obtain, in most experiments we have done, a very small error rate which allow us to avoid an additional step of outliers rejection by estimating the geometric transformation between the pair of images.

In the future we are going to investigate the use of several image features into a single matching process and the matching of non-rigid objects and non-planar scenes.

REFERENCES

Baumberg, A. (2000). Reliable feature matching across widely separated views. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 774–781.

Delponte, E., Isgro, F., Odone, F., and Verri, A. (2006). Svd-matching using sift features. *Graphical Models*, 68:415–431.

Faugeras, O. D. and Berthod, M. (1981). Improving consistency and reducing ambiguity in stochastic labeling: An optimization approach. *IEEE PAMI*, 3(4):412–424.

Hummel, R. A. and Zucker, S. W. (1983). On the fundations of relaxation labeling processes. *IEEE PAMI*, 5(3):267–287.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157. Corfu, Greece.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proc. 13th British Machine Vision Conference*, pages 384–393.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. In *European Conference on Computer Vision (ECCV'2002)*. Copenhag, Denmark.

Mikolajczyk, K. and Schmid, C. (2004). Sacle & affine invariant interest point detectors. *Internationl Journal of Computer Vision*, 60(1):63–86.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Trans on PAMI*, 27(10):1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *Internationl Journal of Computer Vision*, 65(1/2):43–72.

Montesinos, P., Gouet, V., Deriche, R., and Pele, D. (2000). Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18:659–671.

Pelillo, M. and Refice, M. (1994). Learning compatibility coefficients for relaxation labeling processes. *IEEE PAMI*, 16:933–945.

Price, K. E. (1985). Relaxation matching techniques - a comparison. *IEEE PAMI*, 7(5):617–623.

Rosenfeld, A., Hummel, R., and Zucker, S. (1976). Scene labeling by relaxation operations. *IEEE Trans. Systems. Man Cybernetics*, 6:420–433.

Schaffalitzky, F. and Zisserman, A. (2002). Multi-view matching for unordered image sets. In *Proc. 7th European Conference on Computer Vision*, pages 414–431.

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–534.

Tuytelaars, T. and Van Gool, L. (2004). Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85.

Zhang, Z., Deriche, R., Faugeras, O., and Luong, Q.-T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *AI Journal*, 78:87–119.

APPENDIX

Re-writing the Criterion with Matrices

The criterion to be minimized can be written:

$$\begin{aligned} C(x) &= \alpha C_1(x) + (1-\alpha)C_2(x) \\ &= \frac{\alpha}{2n}\sum_{i=1}^{n}\|p_i - q_i\|^2 + \frac{(1-\alpha)m}{m-1}\left[1 - \frac{1}{n}\sum_{i=1}^{n}\|p_i\|^2\right] \\ &= c_1\sum_{i=1}^{n}\|p_i - q_i\|^2 - c_2\sum_{i=1}^{n}\|p_i\|^2 + c_3 \end{aligned}$$

with $c_1 = \frac{\alpha}{2n}$, $c_2 = \frac{(1-\alpha)m}{(m-1)n}$ and $c_3 = nc_2$.

One wants to put C on the form:

$$C([x_1,\ldots,x_n]^T) = \frac{1}{2}\sum_{t=1}^{n}\sum_{p=1}^{n} x_t^T H_{tp} x_p + cte$$

Let remark that the constant is equal to c_3. So one has:

$$\begin{aligned} C(x) &= \sum_{i=1}^{n}(c_1\|x_i - q_i\|^2 - c_2\|x_i\|^2) + c_3 \\ &= \sum_{i=1}^{n}(c_1(x_i - q_i)^T(x_i - q_i) - c_2 x_i^T x_i) + c_3 \\ &= (c_1 - c_2)\underbrace{\sum_{i=1}^{n} x_i^T x_i}_{A} - 2c_1\underbrace{\sum_{i=1}^{n} x_i^T q_i}_{B} + c_1\underbrace{\sum_{i=1}^{n} q_i^T q_i}_{C} + c_3 \end{aligned}$$

The criterion is the weighetd sum of three terms which one notes respectively A, B and C. Let define the following two symbols:

$$\delta_{tp} = \begin{cases} 1 & \text{if} \quad t = p \\ 0 & \text{otherwise} \end{cases}$$

$$\Lambda_{tp} = \begin{cases} 1 & \text{if} \quad a_p \in V_t \\ 0 & \text{otherwise} \end{cases}$$

Then, it is easy to show that:

$$A = \sum_{t=1}^{n}\sum_{p=1}^{n} x_t^T A_{tp} x_p$$

where $\forall \quad t,p \in \{1\ldots n\}$, $A_{tp} = \delta_{tp} I_m$

$$B = \sum_{t=1}^{n}\sum_{p=1}^{n} x_t^T B_{tp} x_p$$

where $\forall \quad t,p \in \{1,\ldots,n\}$, $B_{tp} = \frac{\Lambda_{tp}}{|V_t|} w_{tp} P_{tp}$ and P_{tp} is the matrix of size $m \times m$ containning the conditional probabilities $p_{tp}(k,l)$, and $|V_t| = \#\{V_t\}$.

$$C = \sum_{t=1}^{n}\sum_{p=1}^{n} x_t^T C_{tp} x_p$$

where $\forall \quad t,p \in \{1,\ldots,n\}$, $C_{tp} = \sum_{i=1}^{n}(B_{it}^T B_{ip})$

Finally,

$$C([x_1,\ldots,x_n]^T) = \frac{1}{2}\sum_{t=1}^{n}\sum_{p=1}^{n} x_t^T H_{tp} x_p + c_3$$

with

$\forall \quad t,p \in \{1,\ldots,n\}, \quad H_{tp} = 2(c_1 - c_2)A_{tp} - 4c_1 B_{tp} + 2c_1 C_{tp}$