Matching Local Invariant Features: How Can Contextual Information Help?

Dro Desire Sidibe, Philippe Montesinos, Stefan Janaqi

Matching Local Invariant Features: How Can Contextual Information Help?

Desire Sidibe, Philippe Montesinos and Stefan Janaqi
LGI2P - Ecole des Mines Ales
Parc scientifique G. Besse, 30035 Nîmes Cedex 1, France
{Desire.Sidibe, Philippe.Montesinos, Stefan.Janaqi}@ema.fr

Keywords: image matching, local invariant features.

Abstract – Local invariant features are a powerful tool for finding correspondences between images since they are robust to cluttered background, occlusion and viewpoint changes. However, they suffer the lack of global information and fail to resolve ambiguities that can occur when an image has multiple similar regions. Considering some global information will clearly help to achieve better performances. The question is which information to use and how to use it. While previous approaches use context for description, this paper shows that better results are obtained if contextual information is included in the matching process. We compare two different methods which use context for matching and experiments show that a relaxation based approach gives better results.

1. Introduction

Recently, local invariant features have proven to be very successful in finding corresponding features between different views of a scene. They have been employed in applications such as stereo-vision [1] [12], image retrieval [7], image registration [13], robot localization [6], object recognition [5] [2] and texture recognition [4]. 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. Roughly speaking, matching local invariant features involves three main steps: detecting the interest regions, computing local image descriptors and matching the interest regions using a similarity measure between their descriptors.

An interest region detector is designed to find the same region in different images even if the region is present at different locations and scales. Different methods are proposed in the literature and a good review and comparison is given in [9].

The goal of the description step is to provide, for each region, a vector which captures the most distinctive information within the region. A good descriptor must tolerate small perspective distorsions, illumination changes, image noise and compression. Many different techniques for describing local image regions have been developed and it was shown that the SIFT (Scale Invariant Feature Transform) descriptor performs better than others [8]. This descriptor is based on the gradient distribution in the detected regions and is represented by a 3D histogram of gradient locations and orientations [5].

Once the regions are detected and described, they are matched using a similarity measure between their descriptors.

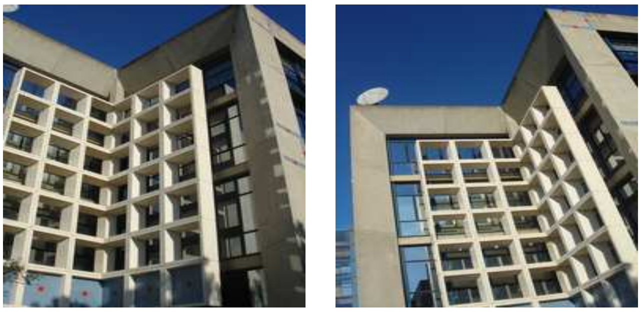

Figure 1: Matching these two images with local features is difficult because of repetitive patterns.

Despite the very good results obtained in different applications, local invariant features are not sufficient to resolve ambiguities that can occur when an image shows multiple similar regions. In the presence of repetitive patterns as in Fig. 1, local features suffer the lack of global information and fail to distinguish between the similar regions.

Differents authors have tried to augment the discrimative power of local feature-based methods by using some *global* or *contextual information*.

One approach is to use contextual information in order to enrich local descriptors. Mortensen, Deng and Shapiro [10] propose a feature vector that includes both local features and global curvilinear information. They use SIFT as local descriptor and shape context [2] as global context descriptor. Similar ideas are used in [13]. While this approach is shown to give better results than SIFT alone, the global context is computed over the entire image and is therefore, sensitive to scale change as well as clutterd background.

Van de Weijer and Schmid [14] add color information to the local shape information. They derive a set of color descriptors which are robust to both photometric and geometric transformations and add them to SIFT feature vector. The combination of SIFT and color lead to better performances as expected, but the obtained gains depend on the application. For a retrieval or a classification task, the combination of color and shape outperforms SIFT alone. But for a matching task, relatively small gains are obtained by adding color to shape information. Moreover, both shape and color descriptors are computed over the small detected regions. Thus, the discrimative power is limited and it will be difficult to distinguish between the similar regions of Fig. 1.

Another approach uses the context in the matching step to resolve ambiguities. Deng et al. [3] propose a framework, called *reinforcement matching*, for including global context into local feature matching. They obtained better results compare with simple matching to nearest neighbour strategy.

Sidibe, Montesinos and Janaqi [11] use contextual information into a relaxation framework and show good performances in comparison with matching to nearest neighbour and SVD-based approaches.

In this paper, we compare these two methods and show that better results are obtained with the relaxation method. In particular, using the robust color descriptors presented in [14] into the relaxation framework described in [11] provides the best results.

2. Using Contextual Information

Local features are not sufficient to resolve ambiguities, because no image descriptor is robust enough to be perfectly discriminant and avoid mismatches. Thus, the idea of using contextual information is to improve matching accuracy by selecting correct matches based on the spatial arrangement of their neighbouring. Local features combined with global relationships convey more information than local features alone. However, global regions are more likely to be sensitive to occlusions and cluttered background. Therefore, contextual information should be defined carrefully.

Let $u = \{u_1, \ldots, u_n\}$ and $v = \{v_1, \ldots, v_m\}$ be two sets of features from two images. Each feature is characterized by a SIFT descriptor. In the next two sections, we briefly described the reinforcement matching and the relaxation matching strategies.

2.1 Reinforcement Matching

As noted by Deng et al. [3], the goal of reinforcement matching is to increase the confidence of a good match between two features if they have a similar spatial arrangement of neighbouring features. First, a cost matrix that contains the Euclidean distance between each pair of features is computed:

$$C = \{c_{ij}\}_{1 \leq i \leq n\ ,\ 1 \leq j \leq m} \tag{1}$$

Then, from this matrix, a fixed fraction (e.g., 20%) of one-to-one best matches are chosen to form *anchor features*. Finally, each detected region is enlarged to form the region context and the cost matrix is updated by combining the initial Euclidean distance with the context score. The context score is obtained by counting, for corresponding bins in the context of two regions, the number of matched anchor features they contain.

$$c'_{ij} = \frac{c_{ij}}{log_{10}(10 + num_{support})} \tag{2}$$

where $num_{support}$ is the number of matched anchor features between the context of the two regions u_i and v_j.

Matches are found using a nearest neighbour with distance ratio (NNDR) strategy, i.e. a feature is matched to its nearest neighbour if that one is much more closer than the second nearest neighbour:

$$d_{ik} = min(D_i) < 0.7\ min(D_i - \{d_{ik}\})$$

where $D_i = \{d_{il}, l = 1, \ldots, m\}$;

2.2 Matching with Relaxation

The relaxation method described by Sidibe, Montesinos and Janaqi [11] is a probabilistic framework which iteratively updates matching probabilities based on a compatibility function. More precisely, let define for each feature u_i a set of initial probabilities:

$$p_i^0 = \{p_i^0(k)\}_{k=1,\ldots,m} \tag{3}$$

$p_i^0(k)$ being the probability that u_i is matched with v_k.

Then, these probabilities are iteratively updated by minimizing a global criterion which takes into account both consistency and ambiguity of the matching. The authors show that the complexity of the method can be drastically reduced if the criterion is written in a convenient way. In particular, they show that the criterion can be written as a quadratic function:

$$C([p_1, \ldots, p_n]^T) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i^T H_{ij} p_j + cte \tag{4}$$

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 contextual information defined by u_i and its neighbour u_j. See [11] for details. The algorithm converges to a local minimum after a reduced number of iterations and for each feature u_i, the feature v_k with highest final probability is retained as its correspondent.

While in they work [11], the authors use normalized cross-correlation to compute contextual information, here we use a more powerful color descriptor as presented in [14]. For each feature u_i and each of its neighbours u_j, we define a circular region, C_{ij}, which diameter is equal to the distance between u_i and u_j. See Fig. 2. We then, compute contextual information by comparing the histograms of *hue* values in both regions C_{ij} and C_{kl}. We use *hue* because it is shown to be robust to photometric and geometric variations [14].

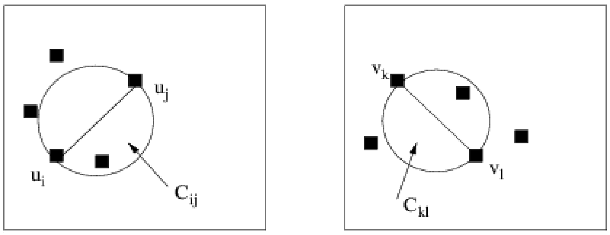

Figure 2: Regions used to compute contextual information in the case of relaxation.

3. Experiments

3.1 Data Set and Procedure

Data Set We compare performances of reinforcement matching and relaxation matching using a publicly available dataset [9] (The dataset is available at http://www.

robots.ox.ac.uk/~vgg/research/affine/). We also use the images presented in Fig. 1 to evaluate the methods in the presence of repetitive patterns. The dataset contains eight sequences of six images each, with growing transformation between the first image and the following ones.

Due to space limitations, we present results obtained for four sequences only. We use two different scene types: structured and textured. And we evaluate three different transformations: viewpoint change, image rotation and scale change. Some of the images are shown in Fig. 3. For each image sequence, the first image is matched to each of the four following ones.

Matching strategies In all experiments, both methods are compare with a standard matching to nearest neihgbour appraoch to see the importance of adding contextual information. Thus, we compare three different matching methods:

- NNDR: nearest neighbour with distance ratio based on SIFT alone [5].
- REINF: reinforcement matching [3].
- RELAX: matching with relaxation [11].

Evaluation criterion We use Harris-Affine regions detector [7] in all experiments. For each image, we keep the 300 detected features with largest cornerness. Then, the features are matched and the matching performance is evaluated based on the number of correct matches obtained for an image pair. We define the matching rate as the ratio between the number of correct matches and the number of detected matches:

$$r = \frac{\#correct\ matches}{\#detected\ matches} \tag{5}$$

Correct matches are detected based on the homographies between the images. A couple of corresponding points (p, p') is said to be a correct match if:

$$\|p' - \mathcal{H}p\| < 5 \tag{6}$$

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

3.2 Results

The comparative results are presented in Table 2, 3, 4 and 5. In the tables, r stands for the matching rate and M for the number of detected matches.

For every sequence, the relaxation based method gives more matches with a matching rate superior or equal to that of the other two methods. For some pairs of images, e.g. the first two images of the *Boat* sequence, RELAX gives as much as twice more correct matches than REINF and NNDR.

For textured scenes (*Bark* and *Wall* sequences), matching to nearest neighbour with SIFT alone gives very good results. Thus a small improvement in performance is observed with REINF and RELAX. On the contrary, the gain is performance obtained by adding contextual information is significant for structured scenes (*Graffiti* and *Boat* sequences).

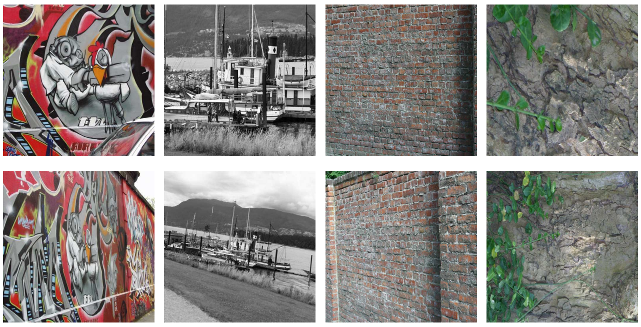

Figure 3: Test images. From left to right: **Graffiti** (viewpoint change, structured scene), **Boat** (scale change + image rotation, structured scene), **Wall** (viewpoint change, textured scene), **Bark** (scale change + image rotation, textured scene).

The matching rate obtained for the *Bark* and *Boat* sequences is almost always close to 1, meaning that the descriptor we use, SIFT, is well suited to rotation and scale changes. For scenes with viewpoint changes (*Graffiti* and *Wall* sequences), the performance of SIFT is very limited as reported in [8]. For this reason, adding contextual information considerably improves the results.

The case of repetitive patterns Matching the images of Fig. 1 is difficult because all the features have almost the same SIFT descriptor. Therefore, matching to nearest neighbour fails in such case and using contextual information becomes necessary. The results obtained for this pair of images are shown in Table 1. As we can see, matching with relaxation outperforms the two other methods. It gives almost twice the number of mathes found by REINF with a higher matching rate. As expected, NNDR gives very poor results.

Method	# detected matches	# correct matches	r
RELAX	38	25	0.66
REINF	16	8	0.50
NNDR	6	3	0.5

Table 1: Results for repetitive patterns.

3.3 Discussion

From the results presented above, we can see that adding contextual information improves the matching results. However, on average, the performance of reinforcement matching is lower than that of matching with relaxation. REINF tries first to increase the matching score of good matches based on the spatial distribution of some *anchor features*. Then, matches are found with a nearest neighbour approach. But if these *anchor features* are not correct, the matching score will not be increased in the right way. Since these *anchor features*are chosen based on the Euclidean distance between SIFT descriptors, they could be incorrect.

The relaxation based approach, increases the probability of a good match based on the configuration of its neighbours. In the method presented in [11], if a match assigned to feature is not *consistent* with those of its neighbours, then this match is discarded. The reason why RELAX performs better than REINF, specially in the case of repetitive patterns, might be the use of color information in the relaxation framework. As noted in [14] and [11], SIFT is based on geometric information alone, so it make sense to add a complementary photometric information.

image	NNDR		REINF		RELAX	
number	M	r	M	r	M	r
2	82	0.96	93	0.96	115	0.92
3	58	0.4	68	0.4	40	0.67
4	23	0.35	24	0.5	24	0.5
5	13	0.08	13	0.08	8	0

Table 2: Results for the **Graffiti** sequence.

image	NNDR		REINF		RELAX	
number	M	r	M	r	M	r
2	70	0.98	91	0.98	150	0.98
3	75	0.95	98	0.95	131	0.98
4	25	0.88	29	0.88	34	0.92
5	19	0.99	22	0.99	24	0.99

Table 3: Results for the **Boat** sequence.

image	NNDR		REINF		RELAX	
number	M	r	M	r	M	r
2	92	0.99	97	0.99	113	0.99
3	41	0.99	51	0.99	78	0.99
4	24	0.87	33	0.85	45	0.92
5	10	0.9	16	0.9	20	0.9

Table 4: Results for the **Wall** sequence.

image	NNDR		REINF		RELAX	
number	M	r	M	r	M	r
2	35	0.97	34	0.97	52	0.98
3	17	0.88	24	0.87	31	0.93
4	3	1	8	1	9	0.9
5	11	0.99	14	0.99	15	0.8

Table 5: Results for the **Bark** sequence.

4. Conclusion

In this paper we have investigated the necessity of using contextual information for matching with local invariant features. Because local features are not sufficient to resolve ambiguities, additional global information is needed. We showed that better results are obtained if contextual information is included in the matching process and we compared two different methods of using context for matching. Experimental results indicate that matching with relaxation performs better than reinforcement matching. The reason being that the former method uses color information which help to distinguish between similar features.

It could be interesting to combine the idea of region context, uses in the reinforcement approach, with the relaxation framework. Moreover, a more powerful descriptor than SIFT could also be useful for applications such as object recognition.

References

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

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans on PAMI*, 24(24):509–522, 2002.

[3] H. Deng, E. N. Mortensen, L. Shapiro, and T. G. Dietterich. Reinforcement matching using region context. In *Proc. "beyond patches" CVPR Workshop*, page 11, 2006.

[4] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Trans on PAMI*, 27(8):1265–1278, 2005.

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

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

[7] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*. Copenhag, Denmark, may 2002.

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

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

[10] E. N. Mortensen, H. Deng, and L. Shapiro. A SIFT descriptor with global context. In *Proc. Computer Vision and Pattern Recognition*, pages 184–190, 2005.

[11] D. Sidibe, P. Montesinos, and S. Janaqi. Fast and robust image matching using contextual information and relaxation. In *Proc. 2nd Int'l. Conf. on Computer Vision Theory and Applications*, pages 68–75, 2007.

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

[13] M. Urschler, J. Bauer, H. Ditt, and H. Bischof. SIFT and shape context for feature-based nonlinear registration of thoracic CT images. In *Proc. CVAMIA, Workshop in conjunction with ECCV'06*, pages 73–84, 2006.

[14] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proc. European Conference on Computer Vision*, pages 334–348, 2006.