Electronic Letters on Computer Vision and Image Analysis 7(1):26-39, 2008

Matching Local Invariant Features with Contextual Information: An Experimental Evaluation.

Desire Sidibe, Philippe Montesinos, Stefan Janaqi

LGI2P - Ecole des Mines Ales, Parc scientifique G. Besse, 30035 Nîmes Cedex 1, France
Received 9th May 2007; revised 25th February 2008; accepted 3rd March 2008

Abstract

The main advantage of using local invariant features is their local character which yields robustness to occlusion and varying background. Therefore, local features have proved to be a powerful tool for finding correspondences between images, and have been employed in many applications. However, the local character limits the descriptive capability of features descriptors, and local features fail to resolve ambiguities that can occur when an image shows 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. Context can be used to enrich the description of the features, or used in the matching step to filter out mismatches.

In this paper, we compare different recent methods which use context for matching and show that better results are obtained if contextual information is used during the matching process. We evaluate the methods in two applications: wide baseline matching and object recognition, and it appears that a relaxation based approach gives the best results.

Key Words: Image matching; Local invariant features; SIFT; Contextual information; Object recognition.

1 Introduction

Recently, local invariant features have proved 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, 21], image retrieval [12], image registration [22], robot localization [10], object recognition [9, 2] and texture recognition [8]. 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 [15].

One of the first methods based on interest points matching is given by Schmid and Mohr [19]. They extract interest points with the Harris detector [6] and use differential invariants as descriptor to match the points. The

Corresponding author is now with the LIRMM laboratory in Montpellier
email: <Desire.Sidibe@lirmm.fr>

Recommended for acceptance by Vittorio Murino

descriptor is computed over relatively small, circular patches around each point. The method is invariant to image rotation and has been extended to color images [16]. Nevertheless, this method fails in the presence of significant transformations, i.e. large viewpoint and scale changes. More recently, there has been a considerable number of works to extend local features and make them invariant to full affine transformations [1, 12, 21, 9, 18, 11]. Among them, it is worth mentioning those based on interest points. Mikolajczyk and Schmid [12, 13] 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 [9] who uses local extrema of Difference-of-Gaussian (DoG) filters as key-points. Similar ideas are used by other authors [1, 18].

The goal of the description step is to provide, for each feature, a vector which captures the most distinctive information within the region around the feature. 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 [14]. This descriptor is based on the gradient distribution in the detected regions and is represented by a 3D histogram of gradient locations and orientations [9].

Once the regions are detected and described, they are matched using a similarity measure between their descriptors. Most of the time, a simple matching to nearest neighbor strategy is used, i.e. a feature in one image is matched to the feature in the other image which is the most closed to it for a given similarity measure.

Despite the very good results obtained in different applications, local feature-based methods are limited in practice by the repeatability of the feature detector and the difficulty of finding enough correct matches in the presence of clutter and large transformations. A simple comparison of the descriptors, for example using Euclidean or Mahalanobis distance, and matching to nearest neighbor 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 popular approach is to estimate the geometric transformation between the pair of images, for example using RANSAC, and use this information to reject inconsistent matches [24]. 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 largely relies on the number of mismatches. Moreover, local invariant features suffer the lack of global information and fail to resolve ambiguities that can occur when an image shows multiple similar regions as in the images of figure 3. In this case, because of repetitive patterns, all the features have almost the same SIFT descriptor and matching to nearest neighbor gives a lot of mismatches.

In many applications, finding a relatively large set of correct matches is crucial. In stereo-vision, a large set of initial correct matches facilitates the estimation of the transformation between two views. Indeed, in the existing literature, RANSAC is used if the portion of mismatches is less than 50% and it is noticed that it fails when the ratio of mismatches is much above this number [10, 3]. In the presence of repetitive patterns, the portion of mismatches might be far greater than 50%. In object recognition applications, because of occlusion and clutter, only a few model features are present in the test image among a large number of non-object features. Therefore, recognition tends to fails because only a few correct matches are found.

In order to reduce to number of mismatches, different authors have tried to augment the descriptive power of local feature-based methods by adding some *global* or *contextual information*.

One approach is to use contextual information in order to enrich local descriptors. Mortensen et al [17] 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 [22]. 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 cluttered backgrounds.

Van de Weijer and Schmid [23] 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 similar regions such as those shown on figure 3.

Another approach uses the context in the matching step to resolve ambiguities. Deng et al [4] propose a framework for including global context into local feature matching called *reinforcement matching*. They obtained better results compare with simple matching to nearest neighbor strategy.

Sidibe et al [20] employ contextual information into a relaxation framework and show good performances in comparison with matching to nearest neighbor and SVD-based approaches.

In this paper, we compare these different methods and show that better results are obtained if contextual information is used during the matching process. In particular, using color information into a relaxation framework 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 information provided by their neighboring. 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 carefully.

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.

2.1 SIFT with Color Information

Van de Weijer and Schmid [23] extend local feature descriptors with color information by adding a color descriptor, K, to the shape descriptor, S:

$$B = \left(\widehat{K}, \lambda\widehat{S}\right) \tag{1}$$

where B is the combined color and shape descriptor, λ is a weighting parameter, and $\widehat{A}$ indicates that the vector A is normalized.

For shape descriptor the authors use SIFT. They try different color descriptors, and show that improvement of the results depend on the application. However, in general, they advice to use the robust *hue* descriptor:

$$hue = artan\left(\frac{O1}{O2}\right) \tag{2}$$

where

$$O1 = \frac{1}{\sqrt{2}}(R - G) \text{ and } O2 = \frac{1}{\sqrt{6}}(R + G - 2B) \tag{3}$$

Furthermore, the histogram of *hue* values is made stable by weighting each sample by its saturation:

$$sat = O1^2 + O2^2$$

2.2 Reinforcement Matching

As noted by Deng et al [4], 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 neighboring features. First, a cost matrix that contains the Euclidean distance between each pair of features is computed:

$$C = \{c_{ij}\}_{1 \leq i \leq n \, , \, 1 \leq j \leq m} \tag{4}$$

Then, from this matrix, a fixed fraction (e.g., 20%) of one-to-one best matches are chosen to form *anchor regions*. 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 matching anchor regions they contain.

$$c'_{ij} = \frac{c_{ij}}{log_{10}(10 + num_{support})} \tag{5}$$

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

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

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

2.3 Matching with Relaxation

The relaxation method described by Sidibe et al [20] is a probabilistic framework which iteratively updates initial probabilities based on a compatibility function. More precisely, let define for each feature u_i a set of initial probabilities:

$$p_i^0 = \{p_i^0(k)\}_{k=1,\ldots,m} \tag{6}$$

$p_i^0(k)$ being the initial 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:

$$C([p_1, \ldots, p_n]^T) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i^T H_{ij} p_j + cte \tag{7}$$

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 provided by the neighbor u_j of u_i. See [20] 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.

For every feature u_i and for each of its neighbors u_j, a circular region C_{ij} which diameter is equal to the distance between u_i and u_j is defined. Then, contextual information is obtained 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 [23].

3 Experiments

The next two subsections present comparative results obtained in two applications: wide baseline matching and object recognition. In each subsection, we first describe the datasets we use and the evaluation criterion for the application. Then, we present the results obtained with different matching methods.

Matching strategies We use Harris-Affine regions detector [12] in all experiments, and the aforementioned methods are compare with a standard matching to nearest neighbor approach to see the importance of adding contextual information. Thus, we compare four different matching methods:

- NNDR: matching to nearest neighbor with distance ratio, based on SIFT alone [9].
- SIFT+COLOR: combined shape and color descriptors [23]. We take the weighting parameter $\lambda = 0.5$ (see Eq. 1).
- REINF: reinforcement matching [4].
- RELAX: matching with relaxation [20].

3.1 Wide Baseline Matching

3.1.1 Data Set

We compare performances of adding color to SIFT, reinforcement matching and relaxation matching using two datasets. The first dataset* contains eight sequences of six images each, with growing transformation between the first image and the following ones [15]. In our experiments, we use four pairs of images from four sequences which represent two different scene types: structured and textured, and three different transformations: viewpoint change, image rotation and scale change. The pairs of images are shown in figure 1. In order to evaluate the methods in the presence of repetitive patterns, we use a second dataset containing four pairs of images presented in figure 3. There are two structured scenes and two textured scenes.

3.1.2 Evaluation criterion

In the case of wide baseline matching, the matching performance is evaluated in terms of *precision* and *recall* of the matching method. These two terms are defined as follows:

$$recall = \frac{\#correct\ matches}{\#correspondences} \tag{8}$$

where $\#correspondences$ stands for the ground truth number of matching regions between the images.

$$precision = \frac{\#correct\ matches}{\#all\ detected\ matches} \tag{9}$$

A couple of corresponding features A and B is assumed to be correct if the *overlap error* is less than 0.5: $\epsilon_S < 0.5$.

ϵ_S measures how well two regions correspond under a known homography H, and is defined by the ratio between the intersection and the union of the regions [14]:

$$\epsilon_S = 1 - \frac{(A \cap H^T B H)}{(A \cup H^T B H)} \tag{10}$$

*The dataset is available at `http://www.robots.ox.ac.uk/~vgg/research/affine/`

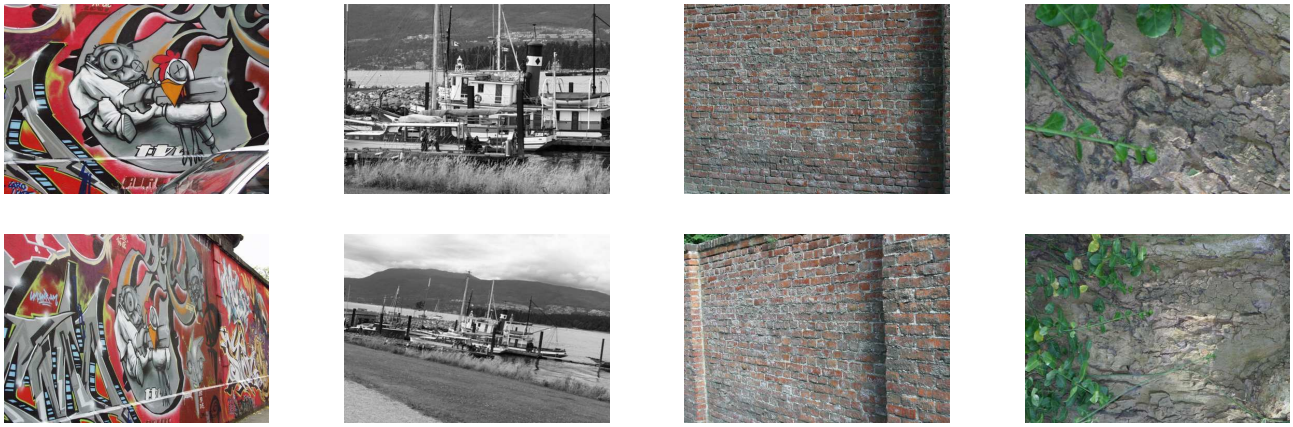

Figure 1: Wide baseline matching test images. From left to right: **Grafitti** (viewpoint change, structured scene), **Boat** (scale change + image rotation, structured scene), **Wall** (viewpoint change, textured scene), **Bark** (scale change + image rotation, textured scene).

3.1.3 Results

Robustness to large transformations

We present the results obtained using the four pairs of images shown in figure 1, with *recall* versus *1-precision* curves. Let note that a perfect matching method would give a recall equal to 1 for any precision. In almost all experiments, we observe that adding color information to SIFT descriptor gives worst results than matching with SIFT alone. This is quite surprising, but is due to the fact that adding color information to the local descriptor increases the ambiguity of matching. More precisely, for the images used in this experiment, except the *Grafitti* and *Boat* pairs, color is not a distinctive information because all the features have the same color. For this reason, SIFT+COLOR gives a reduced number of detected matches, thus reducing the recall.

Despite the relative good performance of SIFT alone, we also observe that substantial gain is obtained with reinforcement and relaxation matching for all pairs of images. In average, RELAX gives approximately 40% higher recall for a precision equal to 0.7. REINF gives approximately 20% higher recall for the same precision. Figure 2 shows the comparative results of the four matching methods. Note that we do not show results with SIFT+COLOR for the *Boat* pair, because it is a pair of greylevel images. The performance of each method depends on the scene type and on the transformation.

• *Scene types:* Considering scene types, we can see that REINF performs well for textured scenes. It gives slightly better results than RELAX for the *Wall* pair of images and similar performances are obtained by both methods for the *Bark* pair. On the other hand, the performance of RELAX is significantly higher than that of REINF for structured scenes (*Grafitti* and *Boat* pairs of images). We can also note that for textured scenes, matching to nearest neighbor with SIFT alone gives very good results, and a small improvement in performance is obtained with REINF and RELAX. On the contrary, the gain is performance obtained by adding contextual information is significant for structured scenes.

• *Transformations:* Regarding the type of transformation, we see that all methods (RELAX, REINF, NNDR and SIFT+COLOR) give higher recall for scale change and rotation (*Boat* and *Bark* pairs of images) compare with the case of viewpoint change (*Grafitti* and *Wall* pairs of images). This can be explained by the fact that the descriptor we use, SIFT, is well suited to rotation and scale changes than to viewpoint changes [14, 20]. For scenes with viewpoint changes, the performance of SIFT is very limited, i.e. the number of detected matches goes down sensitively when the viewpoint change increases . For this reason, adding contextual information considerably improves the results.

Robustness to repetitive patterns

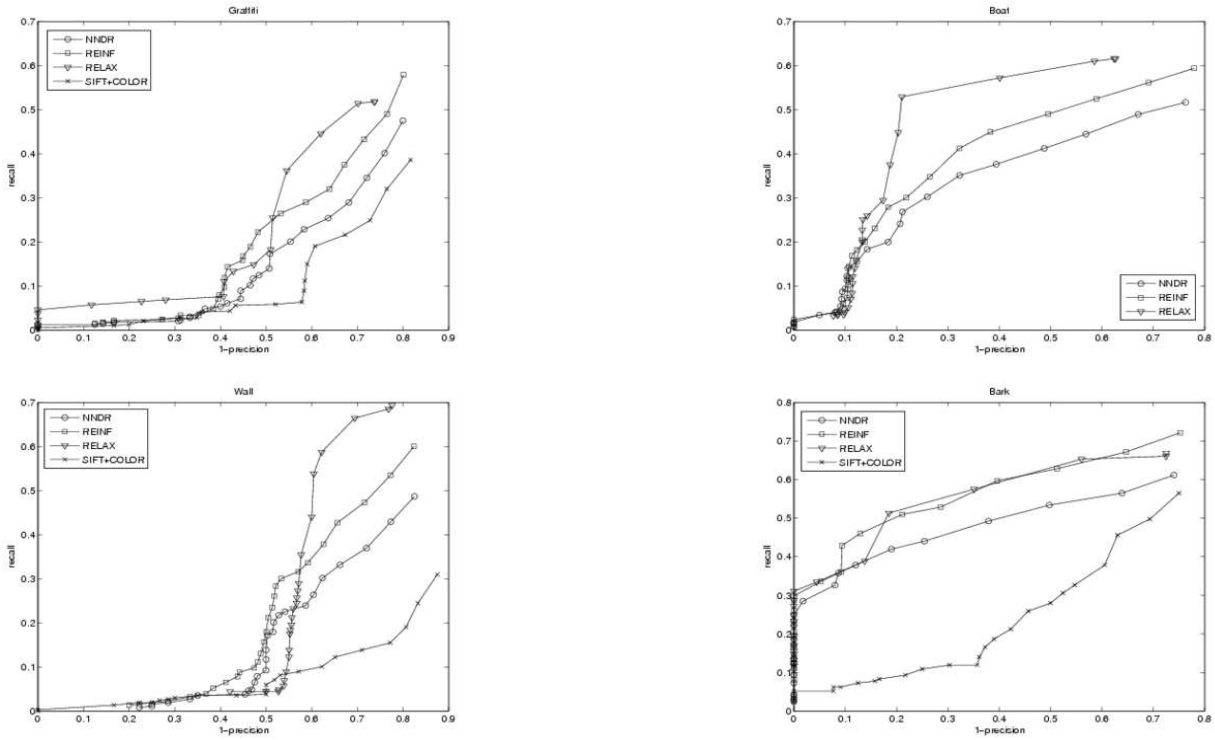

Figure 2: Wide baseline matching results: *recall* versus *1-precision* curves for, from left to right and top to bottom: **Grafitti** pair, **Boat** pair, **Wall** pair and **Bark** pair.

Matching the images of figure 3 is difficult because all the features have almost the same SIFT descriptor. Therefore, matching to nearest neighbor fails and using contextual information becomes necessary. The results obtained for the four pairs of images are shown in table 1. As we can see, matching with relaxation outperforms other methods. It gives almost twice the number of matches found by REINF with a higher precision and a higher recall.

In the case of textured scenes (*Arenas* and *Building* pairs of images), all methods give a high precision. But for structured scenes (*Office* and *Keybord* pairs of images), the precision, i.e. the portion of correct matches, found by NNDR, SIFT+COLOR and REINF is not sufficient to allow the estimation of the geometric transformation between the two views by an algorithm such as RANSAC [10, 3].

3.2 Object Recognition

In this section, we address two recognition applications: object retrieval and object detection. Although these two problems are often mixed up in the literature under the same designation of *object recognition*, there are quite different in nature. The former problem deals with retrieving a given object from a database in which every object is represented by one or more images. The latter, aims to detect and localize a given object in a complex scene which may contain many other objects.

3.2.1 Object retrieval

• *Data Set*
For this application, we use the SOIL-47A dataset which contains 987 images of 47 objects [7]. Each object is imaged 21 times on a black background, at intervals of approximately 9 degrees spanning a range up to ± 90 degrees. The frontal view of all 47 objects constitute the model database, and the remaining 20 images of every

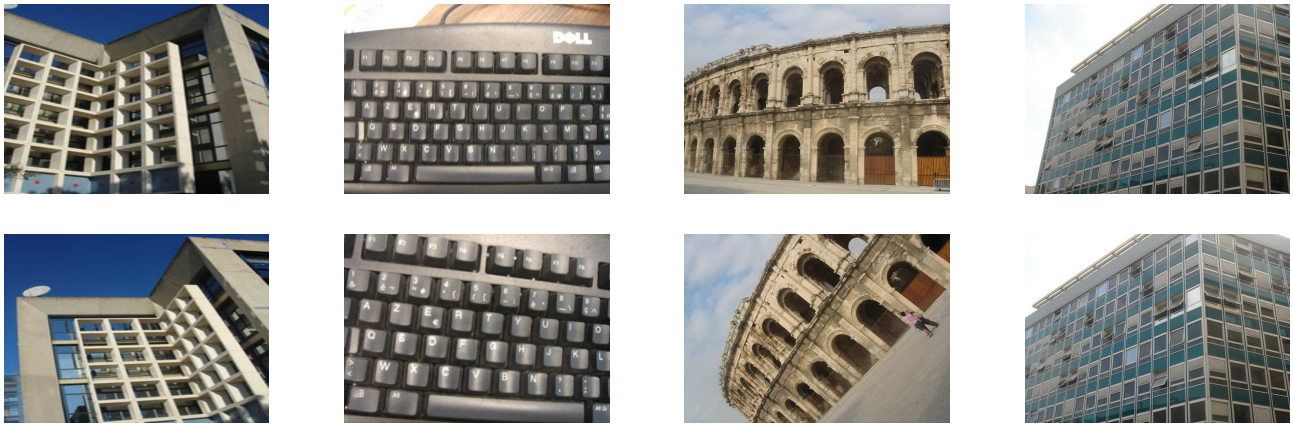

Figure 3: Repetitive patterns test images. From left to right: two structured scenes **Office** and **Keybord**, and two textured scenes **Arenas** and **Building**.

Images	Method	# of matches	precision	recall	time in s
Office	NNDR	6	0.5	0.064	0.165
	SIFT+COLOR	6	0.50	0.064	0.997
	REINF	16	0.5	0.17	0.205
	RELAX	38	0.66	0.53	1.364
Keybord	NNDR	18	0.44	0.1	1.46
	SIFT+COLOR	7	0.42	0.03	2.89
	REINF	18	0.44	0.1	2.17
	RELAX	40	0.66	0.37	3.56
Arenas	NNDR	353	0.94	0.48	2.18
	SIFT+COLOR	170	0.96	0.23	3.66
	REINF	347	0.96	0.47	3.34
	RELAX	568	0.98	0.79	5.22
Building	NNDR	276	0.91	0.44	1.72
	SIFT+COLOR	93	0.82	0.16	3.00
	REINF	300	0.92	0.48	2.61
	RELAX	420	0.98	0.72	5.82

Table 1: Comparison of different algorithms with repetitive patterns (images of figure 3).

Figure 4: Object retrieval experiments: example of objects in the SOIL-47A dataset.

objects are used as test images. It is important to notice that the resolution of the model images is 576x720, while test images have a size 288x360. Figure 4 shows some examples of objects in the SOIL-47A dataset.

• *Evaluation criterion*
For each viewing angle, the performance of a matching method is evaluated by matching all test images viewed under that angle to the model database. Then, for each test image, the model objects are ranked by the number of found matches in decreasing order. An object is correctly retrieved under that angle if the correct model object is among the first k ranks. Finally, recognition performance for this angle is measured as the percentage of rank k correct retrievals.

• *Results*
Results for objects retrieval application are summarized in table 2. As we can see, better performances are obtained when contextual information is used during the matching process. In general, the best performance is obtained for each angle by the relaxation method. Th average correct (rank 1) recognition rate with RELAX is 84.04% for viewing angle in ± 20 degrees and 67.37% for angles in the range ± 60 degrees. When we consider rank 2 and 3, i.e. $k = 3$, the above performances become 97.33% for angles in ± 20 degrees and 80.13% for angles in ± 60 degrees.

As in the case of wide baseline matching experiments, when the viewing angle differs largely from the frontal model view, the recognition rate becomes lower for any method.

3.2.2 Object detection

• *Data Set*
For object detection application, we use a dataset provided by Ferrari et al [5][†]. The dataset is composed of 9 model objects and 23 test images. Some test images contain several objects and in total, the objects appear 43 times in test images. This dataset is a very difficult one since the test images show large viewpoint and scale changes, non-rigid deformations, cluttered background and occlusion up to 89% of the object's surface.

There are 3 planar objects, each modeled by a single view, two objects with curved shapes modeled by 6 views, 3 objects with complex 3D shapes modeled by 8 views, and one frontal view of a 3D object. A single view of each object model is shown in figure 5 and some examples of test images are presented in figure 6. As we can see, figure 6, there is considerable clutter in the test images and the objects appear smaller than in the models, making the matching task a very challenging one.

[†]The dataset is available at `http://www.vision.ee.ethz.ch/~ferrari`

viewing angle in degrees	recognition rate % (rank 1)				recognition rate % (rank 3)			
	NNDR	SIFT+COLOR	RELAX	REINF	NNDR	SIFT+COLOR	RELAX	REINF
-90	0	0	0	0	0.63	12.76	6.38	10.63
-81	0	2.12	2.12	2.12	8.51	23.40	6.38	8.51
-72	2.12	8.51	4.25	2.12	21.27	19.14	23.40	21.27
-63	8.51	8.51	10.63	8.51	25.53	25.63	42.25	23.40
-54	21.27	10.63	31.91	25.53	31.91	25.53	44.68	42.55
-45	42.55	38.29	55.31	48.93	57.44	51.06	63.82	63.82
-36	53.19	29.78	61.70	57.44	76.59	53.19	78.72	78.72
-27	74.46	68.08	76.59	74.46	80.85	72.34	85.10	78.72
-18	61.70	68.08	80.85	72.34	87.23	74.46	95.74	89.36
-9	85.85	51.06	89.36	87.23	89.36	72.34	100	100
+9	80.85	59.57	85.10	82.97	89.36	78.72	100	100
+18	76.59	48.93	80.10	76.59	91.48	70.21	93.61	89.36
+27	63.82	42.55	70.21	70.21	70.21	68.08	85.10	74.46
+36	51.06	40.42	61.70	53.19	72.34	59.57	82.97	78.73
+45	44.68	40.42	57.44	48.93	65.95	53.19	70.21	61.70
+54	48.93	48.93	57.44	57.44	53.19	61.70	61.70	57.44
+63	12.76	21.27	12.76	12.76	25.53	19.14	29.78	25.53
+72	10.63	12.76	10.63	12.76	21.27	29.78	25.53	21.27
+81	6.38	6.38	4.25	6.38	10.63	19.14	21.27	8.51
+90	2.12	0	4.25	2.12	8.51	4.25	10.63	8.51
Total average	37.48	31.64	42.86	40.10	49.89	44.68	56.36	52.12
Average ±60 deg.	58.68	45.56	67.37	62.93	72.16	61.69	80.13	76.23
Average ±20 deg.	76.06	56.91	84.04	79.78	89.36	73.93	97.33	94.68

Table 2: Objects retrieval results using the SOIL-47A dataset. For each viewing angle, the best recognition rate is underlined.

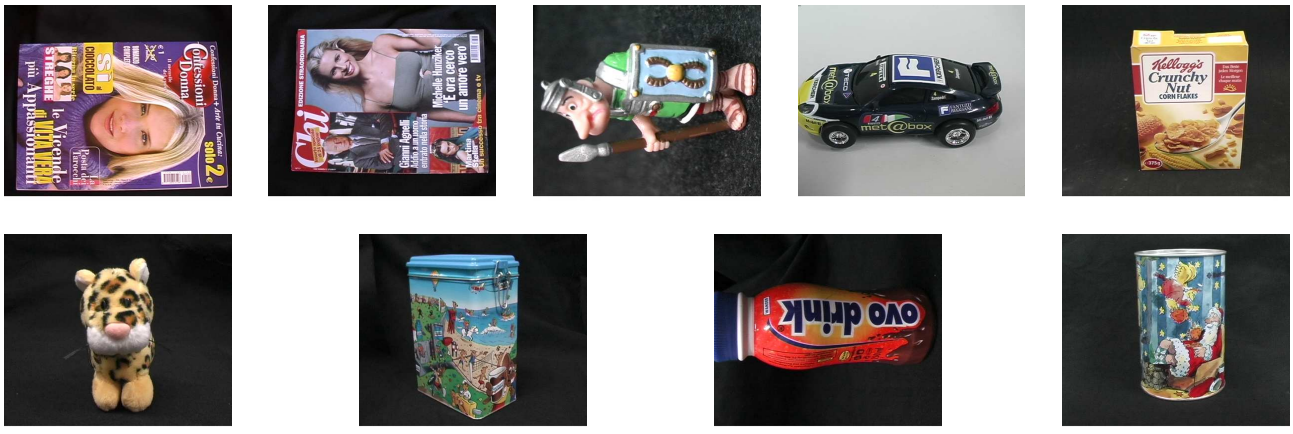

Figure 5: Object detection model objects. From left to right and top to bottom: 3 planar objects, a complex 3D object, 3 objects with complex 3D shapes and 2 objects with curved shapes.

Figure 6: Examples of test images used in object detection experiments.

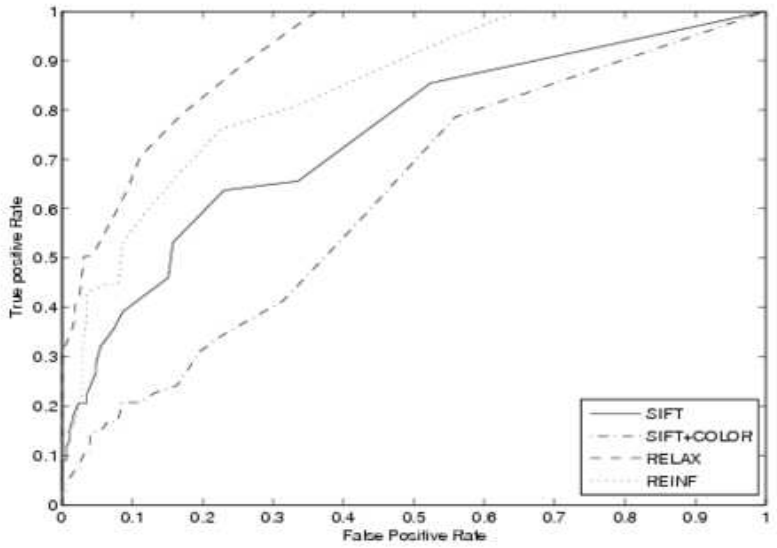

Figure 7: Object detection results.

• *Evaluation criterion*

Performance is evaluated with ROC curves. For every matching strategy, we process all pairs of model-object and test images and we count the number of matched features. An object is detected in a test image if the number of found matches is above a defined threshold. ROC curves are obtained by varying the threshold from 0 to 200 matches as in [5]. Note that if an object is modeled by several views, the number of matches is summed over all its views.

• *Results*

Comparative results are summarized in figure 7. As in the case of wide baseline matching, adding contextual information through the matching process considerably improves the results. RELAX and REINF performs better than NNDR, which itself outperforms COLOR+SIFT. The relaxation based method gives the best results. It achieves a detection rate of 65% with 10% false-positive. For the same false-positive rate, REINF achieves 55% detection rate, NNDR gives 40% detection rate and SIFT+COLOR only achieves 20% detection rate. Nevertheless, none of these methods is satisfactory due to higher level of difficulty posed by the dataset.

3.3 Discussion

From the results presented above, we can see that adding contextual information improves the matching results. 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 neighbor approach. 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 might be incorrect in the presence of clutter or large deformations.

The relaxation based approach, increases the probability of a good match based on the configuration of its neighbors. In the method presented in [20], if a match assigned to a feature is not *consistent* with those of its neighbors, then this match is discarded: i.e. its probability decreases. The reason why RELAX performs better than REINF, specially in the case of repetitive patterns, is certainly the use of color information in the relaxation framework. As noted in [23] and [20], SIFT is based on geometric information alone, so it make sense to add a complementary photometric information which help to distinguish between similar features.

In the case of object retrieval, both REINF and RELAX give better performances than NNDR and SIFT+COLOR. The poor performance obtained by the latter method can be explained by the fact that many objects have very similar color (SOIL-47A dataset), thus increasing the difficulty of retrieving the correct object first. The weak performance in object detection application for all methods is mainly due to the higher level of difficulty posed by the dataset. In mainly cases, only a very small number of correct matches, if any, is obtained when the object is present, which leads to poor recognition performances. However, as the relaxation based method is able to produce more correct matches, it achieves better performance.

We should also notice that Ferrari et al [5] propose a recognition framework based on image exploration which works extremely well on this dataset. The method, first find a set of initial matches and then gradually explores the test image to construct more and more matches. The method achieves 98% detection with 6% false-positive but is computationaly expensive. It takes about 4-5 minutes to process a pair of model and test images [5], which is very slow in comparison with the few seconds needed by the methods presented in this paper.

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 in both wide baseline matching and object recognition applications, 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, using different types of features could also be useful for applications such as object recognition, since it is known that different detectors respond to different types of structures in the image.

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] O. Chum, J. Matas, and S. Obdrzalek. Epipolar geometry from three correspondences. In *Proc. Computer Vision Winter Workshop*, 2003.

[4] 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.

[5] V. Ferrari, T. Tuytelaars, and L. Van-Gool. Simultaneous object recognition and segmentation by image exploration. In *Proc. ECCV*, volume 1, pages 40–54, 2004.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[7] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour-based object recognition algorithms using the SOIL-47 database. In *Proc. of ACCV*, 2002.

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

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

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

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

[12] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Copenhag, Denmark, may 2002.

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

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

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

[22] 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.

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

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