Fusion of Structural and Color Local Descriptors for Enhanced Object Recognition

Pedro Quelhas and Jean-Marc Odobez
IDIAP-Dalle Molle Institute for Perceptual Artificial Inteligence
P.O. Box 592, 1920 Martigny, Switzerland
{pedro.quelhas,odobez}@idiap.ch

Abstract

In this paper we study the behavior of local descriptor object recognition methods with respect to 3D geometric transformations and image resolution variations. As expected performance decreases with accentuated perspective and decrease in resolution. To improve performance and robustness, we propose a scheme to fuse color and gradient local descriptors. This approach is motivated by the discriminative power of color in man-made object recognition. The problem of color feature extraction is addressed as well as the considerations on the fusion process and steps to train such fusion. We used SOIL-47A database for experiments and shown a 7% to 10% relative improvement when compared with state-of-the-art gradient based descriptors.

1. Introduction

The ability to recognize objects in indoor scenes is an essential component of a human-environment interaction understanding system. This leads to the task of object recognition in a situation of low resolution and high geometric variability of the object's appearance.

Object recognition can be based on several aspects of the object's representation in an image: shape [2], color [7, 3], parts organization [14] among others. All these techniques have strong points and weaknesses making them appropriated for different tasks. One technique that has shown successful results consists in the use of collections of local descriptors computed at interest points. It has been used in the past few years to perform recognition tasks such as image retrieval [13, 15] and location identification [11]. Through the introduction of more geometric and scale invariance they have later been adapted for object recognition [10, 11].

These methods are based on local information computed at automatically selected image location and size. No exhaustive scanning is needed. Partial occlusion of the object is handled as far as enough detected locations are left un-occluded so that a positive match is possible. These methods are invariant to viewing conditions like pose, lighting and scale. One important issue with these methods is the matching process which is not easy and relies on local greyscale information [10, 11] that might be ambiguous. In this paper we propose to use local color features to increase the discriminative power of the local descriptors.

Color is known to be a powerful cue for distinguishing and recognizing objects, especially in the case of man-made objects [3]. Color is often used to describe the global content of a full image. Histograms are an example of such a general feature. However, such global features may be difficult to apply to the recognition of objects that occupies only part of the complete image. Besides, there are also difficulties to gather color invariance in an image. To address these issues, Matas et. al. proposed the use of local co-occurrence of color pairs, color bi-modes [7], gathered in small neighborhoods. All locally collected bi-modes are then clustered to derive the final set of main global bi-modes that describe the image. This final step makes this method susceptible to performance loss in the case of heavy background clutter. A second drawback of this approach is that the image spatial distribution of the local bi-mode features is not exploited.

In this article, we analyze the performance of gradient based local descriptor approach [11] with angle and resolution changes on man-made objects using the SOIL-47A database. Surprisingly, the method does not perform as well as reported in the literature [12]. We then propose an extension based on fusing color and greyscale information at the local level. This process results in an improvement of the overall recognition rate.

The method is described in the next Section, while results and discussion are proposed in Section 3.

2. Algorithm Description

In this section we present the different steps of the state-of-the-art local descriptor approach and the proposed color fusion framework. We focus mainly on the details of the color fusion implementation. More detail on the local descriptor methodology can be found in [11].

2.1. Interest Point Detectors and Neighborhood Definition

Local descriptor methods rely on the automatic detection of specific image location p surrounded by a specific image area A_p. The specification and extraction of both the location and area must be reproducible, that is, invariant under geometric and photometric transformations. In this way, areas around a given point will always "cover" the same 3D content (Fig. 1).

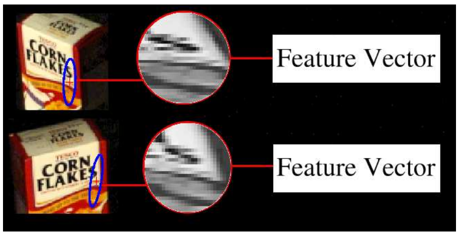

Figure 1. Invariant neighborhood area.

In this paper we exploited the Harris-Affine detector [11, 9]. This choice was motivated by analysis that have shown it to be the most repeatable and stable in the presence of geometric and photometric transforms [12].

2.2. Structural Features

We used steerable filter as structural information descriptors. These descriptors were found to be the best compromise between robustness and dimensionality [12]. Steerable filters are a class of filters in which a filter of arbitrary orientation is synthesized as a linear combination of a set of basis filters [6].

To be used as local structural descriptors, steerable filters are applied to the local area of the image A_p extracted in the previous step (Fig. 1). The responses at a given number of orientations are combined into a set of differential invariants [6] with respect to rotation and illumination.

2.3. Color Features

In [7] Matas defines interest locations for his method as those that have a multi-modal color density function distribution. Local modes where extracted using a mean-shift algorithm. This method is not invariant to scale changes since the width of the kernel in the mean-shift algorithm was set a priori. In our case, we rely on the extracted points p and the associated area A_p. We assume that these regions contain at least a bi-modal color density function distribution.

These color modes are collected using K-means clustering in RGB space. Several experiments where done on a training set of the SOIL-47A database. It was found that estimating reliably the number of modes was difficult, and that the majority of neighborhoods had a bi-modal content. Thus we assume the existence of one only bi-mode at each neighborhood.

We must now use the RGB modes for local description, in an way invariant to possible changes of geometric and lighting conditions. It is well known that these two factors, description and invariance, oppose each other, since increasing invariance results in information loss [8]. In a controlled environment (R, G, B) color values would be the most effective feature. This is however not the case in the presence of illumination changes.

This leads to the choice of an affine invariant illumination model, where we assume that local illumination changes are similar for each mode in the local area. This is a reasonable assumptions in most applications. We adopted the model proposed by Matas [7]. This model makes use of a chromatic color representation, referred to as rg space:

$$r_k\left(R_k, G_k, B_k\right) = \frac{R_k}{R_k + G_k + B_k}, \tag{1}$$

$$g_k\left(R_k, G_k, B_k\right) = \frac{G_k}{R_k + G_k + B_k} \tag{2}$$

For each mode $k = 1, 2$ we compute these features and combined them with the intensity ratio between modes to obtain the local color descriptor c. This ratio is invariant since we assume that both modes undergo the same multiplicative illumination changes. The complete color descriptor is given by:

$$c = \left(r_i, g_i, I_{ij}, r_j, g_j\right), \text{ with } I_{ij} = \left(\frac{R_i + G_i + B_i}{R_j + G_j + B_j}\right) \tag{3}$$

2.4. Correspondence Determination-Validation

Given a query image characterized by its set of interest points and features $q = (p_i^q, v_i^q)_{i=1,\ldots,N_q}$, we look for the object model o, characterized by its list of features $o = (p_i^o, v_i^o)_{i=1,\ldots,N_o}$, which has the largest number of feature matches with q.

Matches are gathered based on a distance between the descriptors. The object feature v_j^o corresponding to a query feature v_i^q is the closest object feature as far as these corresponding features are not to distant ($d^2 < T$). More precisely:

$$\begin{array}{l} j^i = \arg_j \min\left(d^2\left(\mathbf{v_i^q}, \mathbf{v_j^o}\right)\right) \\ \begin{cases} Match\left(\mathbf{v_i^q}\right) = v^o_{j^i},\ if\ d^2\left(\mathbf{v_i^q}, \mathbf{v_i^o}\right) < T \\ Match\left(\mathbf{v_i^q}\right) = 0\ otherwise \end{cases} \end{array} \tag{4}$$

In the case of structural features, the distance used by state-of-the-art methods is the Mahalanobis distance:

$$d_M^2\left(\mathbf{f_i}, \mathbf{f_j}\right) = \left[\mathbf{f_i} - \mathbf{f_j}\right]^T \Lambda^{-1} \left[\mathbf{f_i} - \mathbf{f_j}\right] \tag{5}$$

where Λ is a covariance matrix calculated on a set of training images as explained in [12].

In the case of the color features used in this work we use the following distance [7]:

$$d_C^2\left(\mathbf{c_a}, \mathbf{c_b}\right) = \min\{||c_a - c_b||^2, ||c_a - c_{b'}||^2\} \qquad (6)$$

where $c_{b'}$ represents the color feature vector with the order of the indexes (i, j) switched. This is necessary due to possible variations in the order in which the modes are stored.

It is easy to notice that the simple counting of feature matches can be dominated by false correspondences that must somehow be pruned. This is done here, as proposed by most state-of-the-art methods, by the use of geometric model constraints [11]. A set of correspondences is validated if there exists a valid geometric transformation (homography or epipolar model) between the locations of the points in the query image and their matches in the model image. This solution makes use of the assumption that the object is rigid. Although very effective is is a very computer intensive approach.

2.5. Fusion

Fusion of descriptors can be made in multiple ways and at several stages of a classification process [5]. For the problem at hand we chose feature concatenation as the fusion approach. Two feature vectors f_i and c_i are concatenated into a single feature vector $v_i = (f_i, c_i)$ that is then used for correspondence determination. As described before correspondence determination is based on a distance measure between the descriptors of each point. The distance in the concatenated feature space is then defined as:

$$d_M^2\left(\mathbf{v_i}, \mathbf{v_j}\right) = d_M^2\left(\mathbf{f_i}, \mathbf{f_j}\right) + \alpha d_C^2\left(\mathbf{c_i}, \mathbf{c_j}\right) \qquad (7)$$

where α is a mixing/weighting factor that allows to control the influence of each source on the distance and thus on the final recognition result. The exact value of the mixing parameter α must be trained since its value depends on the unknown importance and reliability of each of the fused features on the definition of an object model.

3. Results

The described method was tested on the SOIL-24A database which is a subset of the SOIL-47A database [1], described in [4]. This subset is composed of 24 images of colorful, planar, household objects; see Fig. 2 for sample images. Object are represented by images of approximately 220x220 pixels at full size. This database was created with the purpose of evaluating the degradation of object recognition methods with respect to the change of viewing angle. In this way we obtain the overall behavior of the system to a possible random positioning of objects. We have sub-sampled the original database to half resolution since objects in human interaction scenes will be smaller than the ones in the original database. For training, images of 4 extra objects not belonging to the SOIL-24A database were selected from the SOIL-47A database. All hyper-parameters were estimated using this training set. In Figure 3 we present the training graph for the parameter α that weights the descriptors' fusion. The optimal value was 0.8 for full resolution images and 1.2 for half resolution images. This curve shows that color influence is more accentuated in the low resolution case. This is illustrated by a greater relative improvement of the performance at the optimal value in the training set. However, at the same time, the performance drops faster as we move away from this value. This phenomenon can be explained by the fact that, the database is known to contain several objects with very similar colors [4]. This results also in an unexpected greater confusion in the matching process (Eq.4) in the low resolution case. Even a small source of confusion can deteriorate the results of this method since only the best match is considered as possible. If color makes several matches have similar distances then the method becomes more prone to errors.

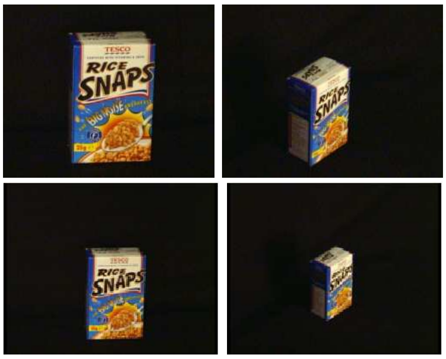

Figure 2. Examples of Soil-24A object images at different angles (0 and 45 degrees) and different resolutions (100% and 50%).

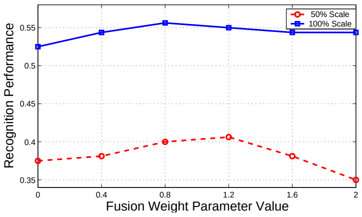

Figure 3. Fusion weight training curves.

Table 1 shows the results of the method when applied to the SOIL-24A database. Structural features produce very good results on near frontal angles but start to break down at high angles. In this case at angles higher than 45 degrees degradation is very high. Unlike reported the structural features did not hold performance above 60% matching performance up to 60 degrees of view angle change [12]; this may be due to higher image resolution of the images in [12] (objects where represented by images with 800x640 pixels).

When applied to the SOIL-24A this color fusion scheme produced overall better results than gradient based features alone. Giving 7% and 10% of relative improvement in relation to the use of only steerable filters for full and half resolution respectively. However, at lower resolutions for some viewing angles the performance is deteriorated. This can be due to the combined influence of the previously mentioned introduced confusion and the fact that the weighting parameter was optimized for the training set.

4. Conclusion

In this contribution we have tested state-of-the-art, gradient based, local descriptor object recognition applied to household objects at reduced resolutions. The results have shown that in low resolution cases local descriptors methods start to have problems dealing with view angle changes.

We have introduced color local descriptors in a fusion framework to aid in the recognition task. It was found that fusing color and gradient features increases the performance of the recognition task. However, the recognition improvment obtained was lower than expected. This is due to the fact that in this database, different images have similar characteristic colors, introducing some confusion in the matching process of local features.

We need to research further the issues raised by this study to understand better the power and limitations of local descriptors that include color. Fusion of color at other levels of the recognition process and methods that allow to take into consideration several possible matches are directions that will be persued.

References

[1] http://www.ee.surrey.ac.uk/research/vssp/demos/colour/soil47/.

[2] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *Proc. ICCV01*, pages 454–463, 2001.

[3] P. Chang and J.Krumm. Object recognition with color cooccurence histograms. In *Proc. IEEE CVPR99*, 1999.

[4] J. M. D. Koubaroulis and J. Kittler. Evaluating colour-based object recognition algorithms using the soil-47. In *Proc. ACCV*, Melbourne, January 2002.

[5] J. K. et. al. On combining classifiers. *IEEE PAMI*, 20(3):226–239, March 1998.

[6] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE PAMI*, 13(9):891–906, 1991.

[7] D. K. J. Matas and J. Kittler. The multi-modal neighborhood signature for modeling object color appearance and applications in object recognition and image retrieval. *Computer Vision and Image Understanding*, 88:1–23, 2002.

Angle	Standard SF		SF Fused with color	
	100%	50%	100%	50%
-90	0.0	0.0	12.5	0.0
-81	4.2	0.0	4.2	8.3
-72	8.3	4.2	8.3	12.5
-63	25.0	12.5	25.0	12.5
-54	33.3	20.8	37.5	16.7
-45	41.7	37.5	58.3	33.3
-36	79.2	62.5	91.7	58.3
-27	91.7	83.3	100	87.5
-18	100	87.5	100	91.7
-9	100	95.8	100	95.8
9	100	95.8	100	95.8
18	100	75.0	100	79.2
27	100	83.3	100	79.2
36	91.7	58.3	91.2	66.7
45	66.7	16.7	75.0	37.5
54	33.3	8.3	33.3	12.5
63	16.7	8.3	16.7	16.7
72	8.3	0.0	8.3	8.3
81	0.0	0.0	4.2	4.2
90	0.0	0.0	4.2	0.0
average	50.0	37.5	53.5	41.1

Table 1. Retrieval performance on the SOIL-24A database, at full and half resolution. (SF stands for Steerable filters)

[8] G. Jan-Mark and R. can den Boomgaard. Color invariance. *IEEE PAMI*, 23(12):1338–1349, December 2001.

[9] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15:415–434, 1997.

[10] D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.

[11] K. Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble, July 2002.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. IEEE CVPR*, June 2003.

[13] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered images sets. In *Proc. ECCV*, pages 414–431.

[14] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. IEEE CVPR98*, pages 45–51, 1998.

[15] T. Tutelaars and L. V. Gool. Content-based image retrieval based on local affine invariant regions. In *Proc. VISUAL99, LNCS 1614*, pages 493–500, 1999.