A LOCAL APPEARANCE CONTEXTUAL DESCRIPTOR FOR OBJECT MATCHING

Xiaozhen Xia, Shuwu Zhang, Wei Liang
Hi-tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{xiaxiaozhen, swzhang, wliang}@hitic.ia.ac.cn

ABSTRACT

We present a novel approach to measuring similarity between objects based on matching local "appearance contextual descriptor". The descriptor has two components: Histogram of Oriented Gradient feature representing local patch appearance and the contextual descriptor capturing not only the spatial distribution of the non-reference patches relative to the reference patch but also the appearance similarities between the reference patch and the non-reference patches in the region. Corresponding patches within two similar objects will have similar contextual descriptors, though the patch appearances may have some difference. We treat recognition in a nearest-neighbor classification framework and match object in regions with no prior learning. We compare our method to commonly used methods and demonstrate its applicability to object detection and recognition.

Index Terms— Object detection and recognition, appearance contextual descriptor, region matching

1. INTRODUCTION

The problem of object matching is a fundamental aspect of many computer vision tasks, including image retrieval, object detection and recognition, action recognition, object tracking, etc. Methods for performing these tasks are usually based on computing local image descriptors, and comparing them using some distance measures.

There is a large number of descriptors which emphasize different image properties like pixel intensities, color, texture, edges, etc. Many of the proposed descriptors use histograms to represent different characteristics of appearance or shape. The SIFT descriptor [6] is represented by a 3D histogram of gradient locations and orientations where the contribution to the location and orientation bins is weighted by the gradient magnitude. An extension of the SIFT descriptor is the GLOH descriptor [7], which uses a log-polar location grid to compute the SIFT descriptor. The shape context [3, 4] describes the coarse distribution of the rest of the shape with respect to a given point on the shape. The HOG descriptor [5] is computed over dense and overlapping grids of spatial blocks, with image gradient orientation features extracted at fixed resolution and gathered into a high-dimensional feature vector. The SURF descriptor [8] uses a Hessian matrix-based measure for the detector and Haar wavelet responses for the descriptor. Other descriptors proposed in the literature include PCA-SIFT [9], CS-LBP [10], OSID [11], and more detailed discussion on local descriptors can be found in [7].

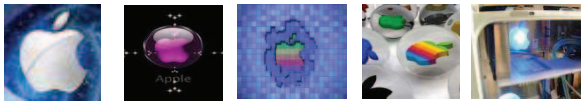

Figure 1: Example of query image and images with various deformation: lens deformation, non-rigid deformation, affine deformation and illumination change.

Many prior work on local features such as SIFT, GLOH, PCA-SIFT and SURF have shown to be fully or partially robust to many of the variations and distortions. While these methods are invariant to scaling, rotations and affine deformation, they do not account for appearance context and can therefore produce ambiguity when matching.

In this paper, we present a new appearance contextual descriptor that combines HOG feature with local contextual descriptor. The HOG feature describes local patch appearance, while the contextual descriptor captures not only the spatial distribution of the non-reference patches relative to the reference patch but also the appearance similarities between the reference patch and the non-reference patches in the region, which helps match similar local regions. Objects are matched in regions with no prior learning, based on a single example image.

The key properties of our descriptor are:

- The feature descriptor is compact and robust across a substantial range of lens deformation, non-rigid deformation, local affine deformation, illumination change, etc.
- Spatial information is naturally encoded into the new descriptor by using the sub-patch configurations to enhance the discriminative power of the descriptor.
- The use of patches as the basic unit captures more meaningful information than individual pixels.

The rest of the paper is organized as follows. In section 2, we provide a detailed description of appearance contextual descriptors. Section 3 gives the procedures of matching appearance contextual descriptors within regions. The experimental results are presented in section 4 and final conclusions are given in section 5.

2. THE APPEARANCE CONTEXTUAL DESCRIPTOR

Fig. 2 illustrates the procedure of generating the appearance contextual descriptor f associated with an image patch p in the image region. Our appearance contextual descriptor consists of two components: HOG descriptor representing patch appearance and local contextual descriptor capturing the appearance similarity between the reference patch and the non-reference patches in the region. Thus, the feature vector is defined as,

$$f = \begin{bmatrix} wa \\ (1-w)c \end{bmatrix} \quad (1)$$

where a is the 36-dimension appearance descriptor, c is a 64-dimension local contextual descriptor, and w is a relative weighting factor. The dimensionality of the appearance contextual descriptor vector is 100.

2.1. Integral histogram of oriented gradients

We use HOG features to represent the appearance of each $m \times m$ patch in an image region. Integral image [12] is applied to efficiently compute the appearance features over a dense regular grid of image patch. For each image, we compute the gradient of each color channel and pick the channel with highest gradient magnitude at each pixel. Then the gradient at each pixel is discretized into one of 9 orientation bins, and an integral image for each bin of the HOG is computed and stored. Finally, the appearance descriptors a of the image patch (typically 8×8) can be computed quickly from those integral images, which contains a 36-dimension histogram vector of concatenating the 9 orientation bins in 2×2 sub-patches.

2.2. The contextual descriptor

The contextual descriptor is computed in a fixed-size image region. For each image, a local region of size $l \times l$ is extracted where the typical choice of l is 48. Suppose there are N patches within the image region, then the contextual descriptor in a reference patch captures the spatial distribution of the remaining patches relative to it and the appearance similarities between the reference patch and the non-reference patches in the region .

For a reference patch p_r with the appearance representation a_r, we compute a contextual descriptor c_r in the image region. The image patch p_r is compared with a larger surrounding image patches centered at p_r . To compare the distance of appearance feature vector between p_r and the remaining patches p_i, $(i \neq r, i = 1,...,N)$ in the region, we use the χ^2 distance defined as,

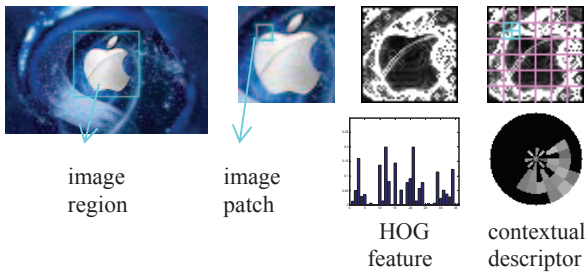

Figure 2: Extracting the local appearance contextual descriptor.

$$d(a_r, a_i) = \frac{1}{2}\sum_{k=1}^{K}\frac{[a_r(k) - a_i(k)]^2}{a_r(k) + a_i(k)} \quad (2)$$

where $a_r(k)$ and $a_i(k)$ denote the K-bin normalized appearance feature vector in p_r and p_i respectively. The resulting distance $d(a_r, a_i)$ is normalized and transformed into a "correlation surface" $S(r,i)$ [2],

$$S(r,i) = \exp(-\frac{1}{W}d(a_r, a_i)) \quad (3)$$

where W is the mean value of the distances between the reference patch p_r and the remaining patches.

The correlation surface is then transformed into log-polar coordinates centered at p_r , and partitioned into 64 bins (16 angle bins, 4 radius bins). The maximal values in all bins form the 64 entries of our local contextual descriptor vector c_r associated with the patch p_r . Finally, this descriptor vector is normalized in order to be insensitive to some noises.

Fig. 3 displays the local contextual descriptor computed at different object locations in two images of the same object. Note that despite the difference in the appearance features between the two images, their local contextual descriptors at corresponding image patches are quite similar.

There are four properties of the appearance contextual descriptor. First, the log-polar representation makes the descriptor invariant to local affine deformation. Second, choosing the maximal correlation value in each bin allows for additional non-rigid deformation. Third, spatial information is naturally encoded into the descriptor by using the sub-patch configurations to enhance the discriminative power of the descriptor. Fourth, the use of patches as the basic unit captures more information than individual pixels.

3. REGION MATCHING

In order to match a query image to a large number of images, we compute the local appearance descriptors a_i and local contextual descriptors c_i densely throughout the images at multiple scales. All the local descriptors in query image form together a global appearance descriptor A and a

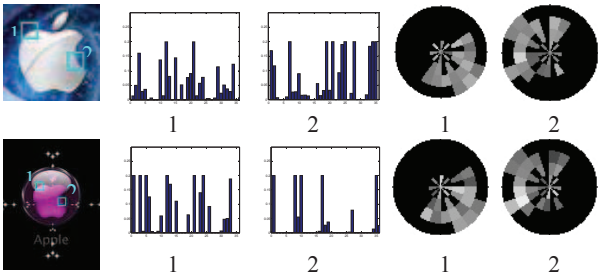

Figure 3: Corresponding appearance features and contextual descriptors. We show patch 1 and patch 2 across tow images of the same object, with their appearance features and contextual descriptors. Despite the difference in appearance features between the two images, their corresponding contextual descriptors are quite similar.

global contextual descriptor C. We would like to find similar A and C in the regions of a collection of images.

To find a good match for the global descriptors of a query image within the regions of other images, we use the efficient approximate-k-nearest-neighbors algorithm and KD-tree implementation of [1]. Given the definition of our feature descriptor in (1) and two global descriptors, A and C, our distance metric is a simple Euclidean distance metric for the global appearance component and a χ^2 distance metric for the global contextual component. The final distance measure value is given by

$$d = wd_A + (1-w)d_C \quad (4)$$

where w is the same weight used in (1), d_A and d_C are distances of global appearance descriptor and global contextual descriptor between regions of the query image and the test image respectively. For the matching results presented here, we use a value of w=0.4.

4. EXPERIMENTS

In this section we experimented with the appearance contextual descriptor, and compare its performance to SIFT and HOG descriptor. Experiment setup is provided in Sec. 4.1. Sec. 4.2 provides some matching results and performance comparisons on our own dataset.

4.1. Experiment setup

Dataset: We evaluated the performance of our descriptor on our own image dataset which is downloaded from the internet. Six image symbols are collected including apple, heart, flower, peace, windows, and pentagram. Each symbol has about 100 images with a variety of deformations such as lens deformation, non-rigid deformation, affine deformation, illumination change, etc. The background images are also collected. **Detector**: Our approach does not require a specific interest point detector. In our experiment, we use a dense regular grid with spacing of 8 pixels. **Evaluation Criterion**: We use recall-precision to evaluate the performance of object matching.

4.2. Object matching in images

We applied the approach presented in the previous section to detect objects of interest in cluttered images. Given a single query image of an object, we densely computed its local appearance contextual descriptors to generate global one. Then we find the appearance contextual descriptor that is maximally similar to that in the regions of detected images. Fig. 4 illustrates some matching results on our own dataset, showing precise matching results despite a substantial range of lens deformation, non-rigid deformation, local affine deformation, illumination change, etc.

In the next experiment, we studied the effect of different parameter configurations on the performance. The performance evaluations varying angle bins (12, 16, 20 and 24) on each dataset are shown in Fig. 5. Table 1 summarizes the average precision scores for each experiment on the 6 symbols. As can be observed, setting the number of angle bins to 12 and 16 give more better results. In general, with more bins, more spatial information can be captured by the descriptor. However, the performance will degrade if there are too many bins relative to the patches due to sensitivity to noise. Also, a high number of bins significantly increases the descriptor's dimension and makes matching more computationally intensive. Fig. 6 plots the average precision score of object matching as a function of the relative weighting factor w used in (1) and (4). As noted earlier, we use a value of w=0.4 in all our results.

We further compared the matching performance of our descriptors with SIFT and HOG descriptors. All the descriptors are extracted over a dense regular grid. Fig. 5 shows the resulting precision-recall curves on the 6 dataset respectively. Table 1 presents the average precision scores for matching performance of different features on the 6 symbols. The results show that our appearance contextual descriptor provided a substantial improvement in accuracy for the six symbols.

5. CONCLUSION

This paper presented a new local appearance contextual descriptor that combines local patch appearance with local appearance context. The descriptor has been evaluated on images with a variety of deformations including lens deformation, non-rigid deformation, local affine deformation, illumination change, etc. Experiments show that the proposed descriptor is very effective in terms of object matching under general image deformations. We believe the proposed descriptor has far reaching implications for many applications in computer vision including object tracking, motion estimation, and action recognition.

Figure 4: Some results on our own dataset. Each row shows matching results using a query image for a specific symbol (Apple, Heart, Flower, Peace, Windows, Pentagram). The first column show the query image for each symbol. Our appearance contextual descriptor is robust across a substantial range of lens deformation, non-rigid deformation, local affine deformation, illumination change, etc.

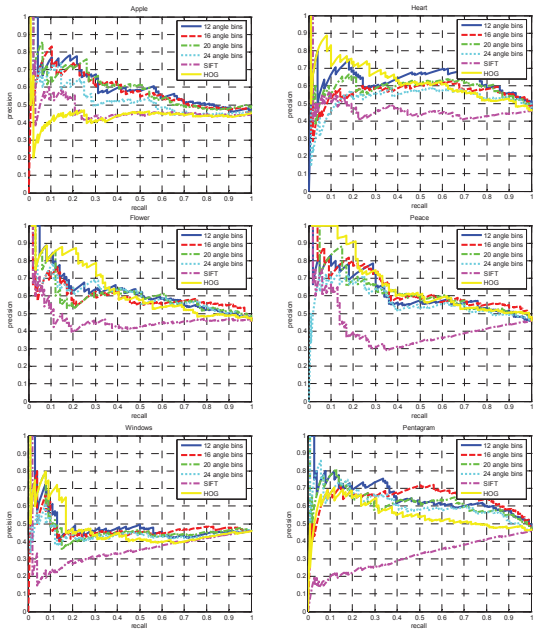

Figure 5: Performance comparison of the appearance contextual descriptor under different parameter configurations (by varying the number of angle bins), and with SIFT/HOG descriptor on the apple, heart, flower, peace, windows and pentagram dataset.

		Apple	Heart	Flower	Peace	Windows	Pentagram
Ours (angle bin)	12	0.6359	0.6724	0.6444	0.6470	0.5350	0.6809
	16	0.6125	0.6365	0.6351	0.6649	0.5237	0.6719
	20	0.6322	0.6481	0.6494	0.6442	0.5307	0.6647
	24	0.5767	0.6052	0.6472	0.5950	0.5196	0.6431
SIFT		0.5248	0.5240	0.5249	0.5251	0.5079	0.4562
HOG		0.5085	0.6553	0.6615	0.6835	0.5322	0.5754

Table 1: Average precision scores for each experiment on the 6 symbols.

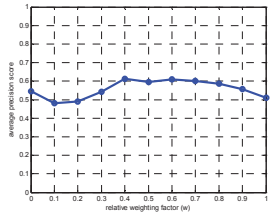

Figure 6: Average precision score of object matching as a function of the relative weighting factor w used in (1) and (4).

6. ACKNOWLEDGEMENTS

The research was supported by National Sciences & Technology Support Program of China (Grant No. 2008BAH21B03, Grant No. 2008BAH26B02 and Grant No. 2008BAH26B03).

7. REFERENCES

[1] J. S. Beis, and D. G. Lowe, "Shape Indexing using Approximate Nearest-Neighbor Search in High-Dimensional Spaces," *CVPR*, pp. 1000-1006, 1997.
[2] E. Shechtman, and M. Irani, "Matching Local Self-Similarities across Images and Videos," *ICCV*, 14(1):1-8, 2007.
[3] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *PAMI*, 24(24):509-522, 2002.
[4] G. Mori, S. Belongie, and J. Malik, "Efficient Shape Matching Using Shape Contexts," *PAMI*, 27(11):1832-1837, 2005.
[5] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," CVPR, pp. 886-893, 2005.
[6] D. G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, 60(2):91-110, 2004.
[7] K. Mikolajczyk, and C. Schmid. "A Performance Evaluation of Local Descriptors," *PAMI*, 27(10):1615-1630, 2005.
[8] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," *ECCV*, pp. 404-417, 2008.
[9] Y. Ke, and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *CVPR*, pp. 506-513, 2004.
[10] M. Heikkila, M. Pietikainen, and C. Schimid, "Description of interest regions with local binary patterns," Pattern Recognition, 42(3):425-436, 2009.
[11] F. Tang, S. H. Lim, and N. L. Chang, "A Novel Feature Descriptor Invariant to Complex Brightness Changes," *CVPR*, 2009.
[12] P. Viola, and M. Jones, "Rapid Object Detection Using A Boosted Cascade of Simple Features," *CVPR*, pp. 511-518, 2001.