ORDINAL PYRAMID POOLING FOR ROTATION INVARIANT OBJECT RECOGNITION

Guoli Wang, Bin Fan and Chunhong Pan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{glwang, bfan and chpan}@nlpr.ia.ac.cn

ABSTRACT

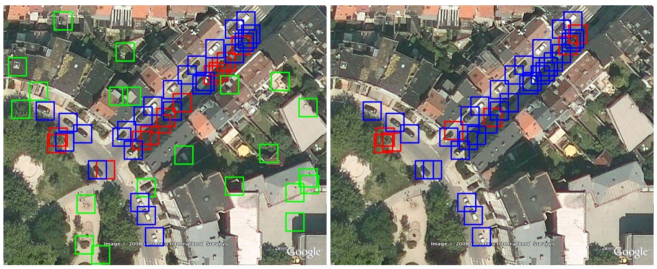
Local feature descriptor plays a fundamental role in many visual tasks, and its rotation invariance is a key issue for many recognition and detection problems. This paper proposes a novel rotation invariant descriptor by ordinal pyramid pooling of local Fourier transform features based on their radial gradient orientations. Since both the low-level feature and pooling strategy are rotation invariant, the obtained descriptor is rotation invariant by nature. Pooling based on orders of gradient orientations is not only invariant to in-plane rotation, but also encodes gradient orientation information into descriptor as well as spatial information to some extent. Moreover, these information is enhanced by the proposed pyramid pooling structure. Therefore, our method is naturally rotation invariant and has strong discriminative ability. Experimental results on the aerial car dataset demonstrate the effectiveness of our descriptor.

Index Terms— Local feature descriptor, Rotation invariant, Orders of radial gradient orientations, Ordinal pyramid

1. INTRODUCTION

Local feature descriptors computed from image patches have been widely studied in recent years. They have been used in a variety of visual tasks, such as human detection [1], object classification [2] and tracking [3]. These tasks require local feature descriptors with strong discriminative abilities and good robustness to illumination changes, background clusters, partial occlusions and so on.

Perhaps one of the most popular local feature descriptors is HOG (Histogram of Oriented Gradients) [4, 5]. It divides the image patch into small cells and aggregates histogram of oriented gradients in each cell. Its subregion and histogram accumulation structure provides more spatial information and more robust presentation. However, HOG is not invariant to in-plane rotation, which limits its application for detecting rotated objects, such as cars in remote sensing images. Fig. 1(a) shows an example of car detection by HOG, where true positives are detected along with many error detections.



(a) HOG (b) OPP

Fig. 1. Detection results of the aerial car candidates by (a) HOG and (b) our proposed descriptor. True positives are indicated by blue boxes while false positives are marked with green boxes, and red boxes indicate false negatives.

In literature there are three prominent ways to achieve rotation invariant descriptors. The first one estimates a dominant orientation of the image patch and calculates descriptor relative to this dominant orientation, e.g. SIFT (Scale Invariant Feature Transform) [6, 7] and SURF (Speeded Up Robust Features) [8]. However, Fan et al. [9, 10] experimentally proved that dominant orientation estimation is not exact and orientation assignment error is the major cause for bad recognition and matching results. The second way is to calculate local descriptor in polar coordinates, such as spin image, RIFT (Rotation Invariant Feature Transform) [11] and RIFF (Rotation-Invariant Fast Feature) [3], but this kind of methods lacks discrimination as the loss of spatial information due to their ring shaped pooling schemes. The third technique is a recently proposed one, called Fourier HOG [12] which achieves rotation invariance by representing features in Fourier space. Since it loses partial phase information, its discriminative ability could be further improved.

In this paper, we propose a novel method for local descriptor construction which pools local Fourier transform features [12] based on orders of their radial gradient orientations in a pyramid manner. Fig. 2 shows that the radial gradient orientations of the same parts of an object at different rotating angles are almost identical and the radial gradient orien-

This work was supported by the National Natural Science Foundation of China under Grants 91338202, 61203277, 61272394, 91438105 and 61305049.

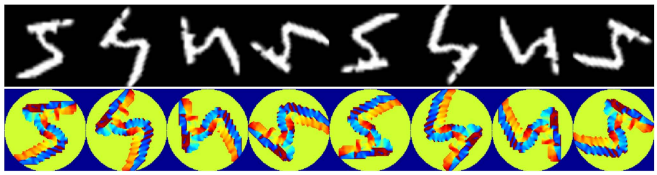


Fig. 2. An illustration of the rotation invariance of radial gradient orientation. A handwritten digit image (140×140) is rotated by 8 different orientations (Top). Their radial gradient orientations are indicated by different colors (Bottom).

tations in local region are similar. Therefore, sorting radial gradient orientations of sample points is rotation invariant. A pooling scheme based on their orders not only maintains the rotation invariance, but also encodes local regional information into the descriptor and improves its discriminative ability. Moreover, inspired by the spatial pyramid [13], we propose to partition orders of radial gradient orientations into increasing groups as a pyramid and pool local Fourier transform features inside each groups. The pooled representation in each pyramid level is concatenated as the final descriptor. Fig.1(b) shows that our method can recognize more true positive with less error detections. Our main contributions include:

- Local features are pooled by orders of their radial gradient orientations, which are invariant to rotation. Such a pooling strategy can encode both gradient orientation and spatial information to some extent.
- A ordinal pyramid framework is proposed to further improve the discriminative ability.

The rest of this paper is organized as follows: Section 2 describes how to construct the proposed descriptor, followed by experiments in Section 3. Finally, we conclude this paper in Section 4.

2. DESCRIPTOR CONSTRUCTION

In this part, we elaborate construction of the proposed descriptor. In section 2.1, we first analyze the rotation invariance of orders of radial gradient orientations, and then we show how to obtain more spatial and ordinal information in a ordinal pyramid framework. Fourier transform feature is used as low-level feature for pooling in our method and Section 2.2 gives a brief introduction of it.

2.1. Ordinal pyramid pooling

The proposed descriptor is based on radial gradient orientations. The radial gradient is firstly introduced in RIFT [11] and the proof of its rotation invariance is given in RIFF [3]. Suppose O is the centre of an image patch and P is a sample point in the image patch, we establish a local x-y coordinate

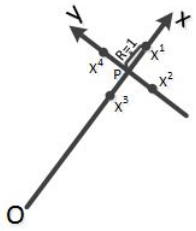


Fig. 3. The radial local coordinate system used for gradient computation.

system of P as shown in Fig.3, where $\overrightarrow{OP}$ is defined as the positive x-axis. In this coordinate system, the gradient of P can be computed as:

$$dx(P) = I(x^1(P)) - I(x^3(P)), \tag{1}$$

$$dy(P) = I(x^4(P)) - I(x^2(P)). \tag{2}$$

We call this gradient as radial gradient, and its orientation is computed as:

$$\theta(P) = \tan^{-1}\left(\frac{dy(P)}{dx(P)}\right). \tag{3}$$

Obviously, radial gradient is rotation invariant as this local coordinate system is constructed independent of image rotation. Fig. 2 gives an example, in which a handwritten digit image is rotated by 8 different orientations (Fig. 2 (Top)). Their radial gradient orientations are calculated and shown in Fig. 2 (Bottom). We can observe that: (1) the radial gradient orientations on the same parts of different rotated images are almost same; (2) the radial gradient orientations in local region are similar to each other. According to the first observation, when we sort sample points in the image patch according to their radial gradient orientations, the sorted list of samples would not be affected by rotation. Meanwhile, because of the second observation, sorting by gradient orientations tends to preserve spatial adjacency. Consequently, on the one hand, partitioning sample points into groups by their orders of gradient orientations could encode ordinal information of gradient orientation while maintaining rotation invariance. On the other hand, pooling in the ordered groups provides stable local region information as sample points in local region tend to be partitioned into same groups.

Suppose $\{X_1, X_2, \cdots, X_N\}$ are N sample points in an image patch, and $\theta(X_i)$ is the radial gradient orientation of X_i. Sorting these sample points according to their orientations, we can obtain a set of sorted sample points:

$$\left\{ \begin{array}{c} X_{f(1)}, X_{f(2)}, \cdots, X_{f(N)} : \\ \theta(X_{f(1)}) \le \theta(X_{f(2)}) \le \cdots \le \theta(X_{f(N)}) \end{array} \right\},$$

where $f(1), f(2), \cdots, f(N)$ is a permutation of $1, 2, \cdots, N$.

Once the sample points are sorted by their gradient orientations, we partition them into k groups equally and local features are pooled together in these groups. Inspired by the spatial pyramid [13], we consider partitioning sample points and pooling their low-level features by a pyramid strategy, which we call ordinal pyramid pooling. More specifically, we partition the sorted samples in multiple levels in a pyramid framework and $k = 2^{level-1}$ groups are obtained in each level. Then we use the sum pooling operation, i.e., all the low-level features of the sample points in each group are summed together to obtain a pooled vector, and those pooled vectors concatenated together in each level are normalized to unit length. Finally, the normalized vector of each pyramid level is concatenated as the final descriptor. This pyramid pooling method not only provides more mutual information about spatial layout and gradient orientation distribution, but also improves robustness to shift of orientation order. For example, in case of different background, the orientations of sample points in background would affect the orientation orders of sample points in object in interest. By the pyramid pooling, the redundant information representation can reduce this influence and improve the discriminative ability as a result.

2.2. Local features in Fourier space

Fourier HOG [12] is a rotation invariant local descriptor for image patches. Although losing partial phase information, it has strong discriminative ability because it is represented by the continuous distribution and is rotation invariant theoretically. In this paper, we use the Fourier HOG to describe a 25×25 local region of each sample point as its low-level feature. Due to the space limit, readers are referred to [12] for a detailed description of Fourier HOG, here we describe some modifications that we adapt so as to make it flexible in our method: (1) we reduce the regional feature sample radius from 6 to 4; (2) we give up the coupling features between different radii since we only need a local representation of each sample point; (3) we convert some dimensions of Fourier HOG vector of each sample point into double dimensions as they may have negative values, which are not suitable for pooling. In this case, a value R is converted into [max(R,0) max(-R,0)]. Finally, each sample point has a 143 dimensional Fourier HOG feature.

Algorithm 1 gives the pseudo code of our descriptor construction. As our descriptor pools local Fourier transform features in ordinal pyramid pooling framework, we call our method as *OPP*.

Algorithm 1 Ordinal Pyramid Pooling Descriptor Construction

Input: Image patch I
Output: Patch descriptor f

1: Calculate Fourier HOG feature F_i and radial gradient orientation for each sample point X_i
2: Sort sample points by their radial gradient orientations
3: **for all** pyramid level l **do**
4: Partition the ordered samples into $k = 2^{l-1}$ groups
5: Pool local feature F_i of sample points in each group
6: Concatenate the pooled vectors to form a feature vector $fp(l)$ of level l and normalize $fp(l)$ to unit length
7: **end for**
8: Concatenate all $fp(l)$ to form the final descriptor f

3. EXPERIMENTS

To evaluate the effectiveness of our proposed method, here we conduct experiment on aerial car recognition, and compare its performance with the state of the art.

3.1. Dataset description

The aerial car dataset [14] is used in our experiment. It consists of 30 satellite images with 1319 labeled cars. The cars are rotated arbitrarily but only annotated with axis-aligned bounding boxes. The dataset provides candidates for detection with fixed window size of 45×45 pixels. 15 even named images containing 9339 candidates are chosen for train and other 15 images with 9850 windows are used for test.

3.2. Evaluation criterion

Performance of different methods are measured by average precision recall (AP), which is widely used in PASCAL VOC [15]. Let $Area(Target)$ be pixels in the bounding box and let $Area(GroundTruth)$ be the ground truth set of pixels where the car is actually located. A detection window is labeled as a car if the $Area(Target)$ and the $Area(GroundTruth)$ satisfy the following condition:

$$\frac{Area(Target \cap GroundTruth)}{Area(Target \cup GroundTruth)} \geq 0.2, \qquad (4)$$

$Precision = \frac{TP}{TP+FP}$ is the number of true positives divided by the total number of detected positives and $Recall = \frac{TP}{TP+FN}$ is the number of true positives divided by the total number of groundtruth positives. They are calculated in a ranked threshold sequence, and can be plotted as a precision-recall curve. At eleven equally spaced recall levels $t = \{0, 0.1, \cdots, 1\}$, the precision at recall level t_i is calculated as the maximum in all precisions whose corresponding recalls exceed t_i. The mean precision of all these recall levels is defined as AP.

3.3. Evaluated methods

HOG [4] is the common descriptor used for object recognition, hence we use it as baseline. Its implementation [16] is 8×8 in pixel cells and 9 orientation bins in each cell. Fourier HOG [12] is a recently proposed descriptor for rotation invariant object recognition. We use the implementation supplied

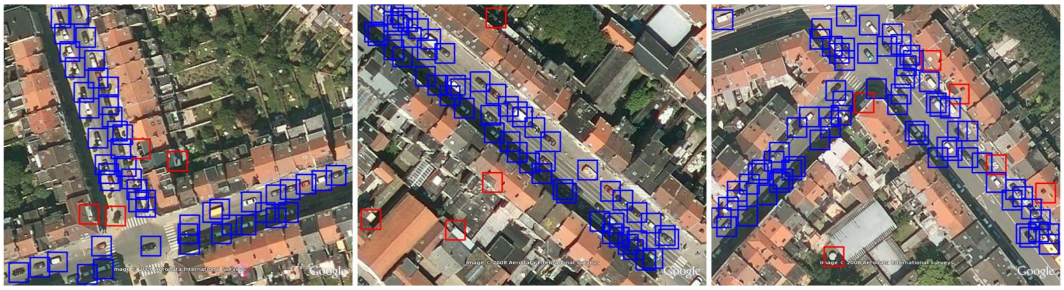


Fig. 4. Detection results of the aerial car candidates by OPP (true positives in blue, false positives in red)

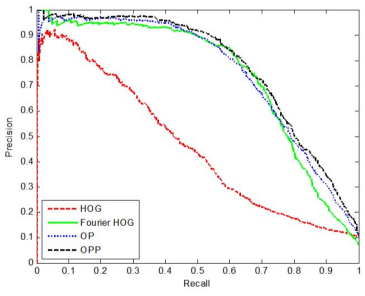


Fig. 5. Precision-Recall curve by different methods on the aerial car dataset.

Table 1. The average precision of different methods on the aerial car dataset.

Methods	HOG	Fourier HOG	OP	OPP
AP(%)	46.96	72.58	74.00	**75.94**

by authors. Two methods (OP and OPP) of the proposed descriptor are evaluated. OP is our proposed ordinal pooling descriptor without using the pyramid structure. In this implementation, we divide sample points in an image patch into 16 groups according to their gradient orientation orders, combined with 143 dimensional Fourier HOG feature, OP finally gets a 2288 dimensional descriptor used for object recognition. OPP is the proposed ordinal pyramid pooling descriptor, in which a 5-level pyramid is used, resulting a 4433 dimensional descriptor used for object recognition. The extracted features are fed into SVM classifier with linear kernel [17], in which the parameter C is selected via five-fold cross validation.

3.4. Results and analysis

The precision-recall curves for HOG, Fourier HOG, OP and OPP are depicted in Fig. 5, and the average precisions are listed in Table 1. As can be seen, HOG performs the worst in this task as it is sensitive to image rotation. Owing to the rotation invariant property, Fourier HOG achieves a much improved performance. It is clear that both OP and OPP outperform Fourier HOG, demonstrating the effectiveness of the proposed pooling strategy. In particular, Fig. 5 shows that OP and OPP get higher precisions than Fourier HOG when the higher recall rate (≥ 0.8) is required. This means that OP and OPP detect less false negatives when more true positives are demanded. OPP further improves OP about 2%, and achieves the best performance in this dataset. This indicates the importance of the proposed ordinal pyramid, which could encode coarse to fine ordinal information of the gradient orientations. To sum up, both pooling based on the ordered orientations and the ordinal pyramid are critical for improving the discriminative ability. Both of them contribute to the good performance of our method. Some detection results are shown in Fig. 4, where most of cars are detected with a few false positives. It is noted that OPP also detects some cars in shadow or poor centralized, which are difficult to be detected.

4. CONCLUSIONS

This work proposes a novel method for rotation invariant descriptor construction. The main contributions of our work are from two aspects. Firstly, pooling by orders of radial gradient orientations is not only invariant to in-plane rotation, but also encodes gradient orientation information into descriptor as well as spatial information to some extent. Secondly, the ordinal pyramid framework which includes coarse to fine gradient orientation ordinal and spatial information further improves the discriminative ability of our proposed descriptor. Experimental results show that OPP is suited to recognize rotated objects and can achieve better performance.

The potential improvement of the discriminative ability remains to be investigated in our future work. For example, radial gradient orientations of background sample points may affect the stability of orientation orders, and this may lead to wrong recognition results, hence more discriminative feature could be considered to form robust and rotation invariant orders for samples. In addition, more spatial information could be obtained by the ring shaped structure, and concatenating feature vectors by ordinal pyramid pooling in rings may improve the discriminative ability.

5. REFERENCES

[1] Q. Zhu, M.C. Yeh, K.T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 1491–1498.

[2] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[3] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Unified real-time tracking and recognition with rotation-invariant fast features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 934–941.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[6] D.G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[7] W. Zhang, X. Sun, K. Fu, C. Wang, and H. Wang, "Object detection in high-resolution remote sensing images using rotation invariant parts based model," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 74–78, 2014.

[8] H. Bay, T. Tuytelaars, and L. Van-Gool, "Surf: Speeded up robust features," in *Computer Vision-ECCV 2006*, pp. 404–417. Springer, 2006.

[9] B. Fan, F. Wu, and Z. Hu, "Rotationally invariant descriptors using intensity order pooling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 2031–2045, 2012.

[10] B. Fan, F. Wu, and Z. Hu, "Aggregating gradient distributions into intensity orders: A novel local image descriptor," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2377–2384.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1265–1278, 2005.

[12] K. Liu, H. Skibbe, T. Schmidt, T. Blein, K. Palme, T. Brox, and O. Ronneberger, "Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 342–364, 2014.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[14] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Computer Vision-ECCV 2008*, pp. 30–43. Springer, 2008.

[15] M. Everingham, L. Van-Gool, C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[16] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1469–1472.

[17] C.C Chang and C.J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.