Rotation-invariant Histograms of Oriented Gradients for Local Patch Robust Representation

Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki
Graduate School of System Informatics, Kobe University, Kobe, 657-8501, Japan
{luozhaojie, ianchen}@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

***Abstract*—Our research focuses on the question of feature descriptors for robust effective computing, presenting a novel feature representation method– rotation-invariant histograms of oriented gradients (Ri-HOG). Most of the existing HOG techniques are computed on a dense grid of uniformly-spaced cells and use overlapping local contrast of rectangular blocks for normalization. However, we adopt annular spatial bins type cells and apply radial gradient transform (RGT) to attain gradient binning invariance for feature descriptors. In such way, it significantly enhances HOG with respect to rotation-invariant ability and feature descripting accuracy. In experiments, the proposed method adopts object recognition as a test case and it is evaluated on PASCAL VOC 2007 dataset. The experimental results demonstrate that the proposed method is much more efficient than the existing methods.**

I. Introduction

Representing salient patches for images in a way that is capable of invariant representation is a significant challenge in Computer Vision, as it lies at the core of many popular classification applications, such as object recognition, affective computing, image retrieval *etc*. To invariantly describe the non-linear information of the images, well-known local feature descriptors, such as SIFT [1], SURF [2], *etc.*, typically apply a set of hand-crafted filters and aggregate their responses within pre-defined regions of image patches. The extent, location and shape of these regions define the aggregating configuration of the descriptor. Recently, some works show that optimizing this configuration can result in fairly large performance improvement [3], [4]. Although interesting solutions have been presented, these approaches are either based on hand-crafted representations [3]; or still require many complex and high-cost parameter transformations (*e.g.*, in [4], it relies on a non-analytical objective that is difficult to optimize). Therefore, although these feature descriptors have claimed to invariant representation, they still have various challenging problems.

In this paper, we propose a novel feature representation method, called rotation-invariant histograms of oriented gradients (Ri-HOG), for the image local features on invariant representation. For robustness and speed, we carry out a detailed study of the effects of various implementation choices in descriptor performance. We subdivide the local patch into annular spatial bins, to achieve spatial binning invariance. Besides, we apply radial gradient to attain gradient binning invariance, which is derived from the theory of polar coordinate. By doing so, the proposed method can significantly enhance the features descriptors in regard to invariant representation ability and feature descripting accuracy.

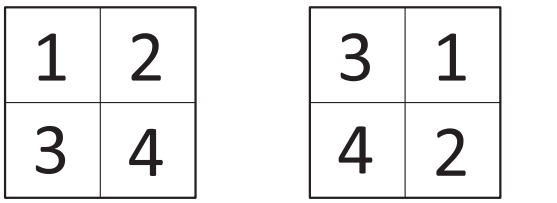
The main contribution of this paper is to present an image representation method, called Ri-HOG, which is an appropriate similarity measure that can remain invariant with respect to image transformations. The proposed method is very significant, as it is very important to those with closely related research interests mentioned above. Moreover, based on the linear SVMs, the proposed feature is applied to a very popular classification application–object recognition as a test case in our experiments, which are valid on PASCAL VOC 2007 [5] database. As a result, its performance reaches the state-of-the-art level.

In the remainder of this paper, we describe the proposed method and additionally have a look at the related works in Sect. II. Sect. III gives the detailed stages of process in experimental evaluations and conclusions are drawn in Sect. IV.

II. Proposed Method

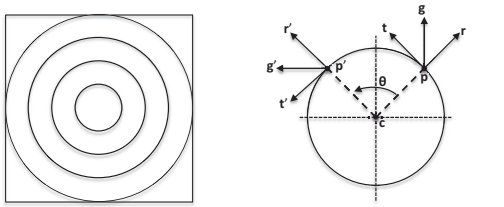
A. Background and problems:

HOG are feature descriptors, which are computed on a dense grid of uniformly-spaced cells and use overlapping local contrast normalization for improved accuracy. These features are set based on *cells* and *blocks* representation system is widely used in classification applications, especially human detection. The describing ability of HOG features set outperforms many existing features [6], however, its robustness against image rotation are not satisfactory. Here one direct evidence is that the HOG feature is seldom applied to object tracking or image retrieval successfully. Giving a more scientific reason, see Fig. 1 for an example. Supposing Fig. 1(a) is an image with HOG block size, there are 4 cells in the block. Fig. 1(b) is an image of Fig. 1(a) after making a quarter turn. HOG features are extracted from the two images individually. If the histogram of oriented gradients obtained from the regions 1, 2, 3, and 4 are severally denoted as x_1, x_2, x_3, x_4, then, the HOG features extracted from Fig. 1(a) and Fig. 1(b) are (x_1, x_2, x_3, x_4) and (x_3, x_1, x_4, x_2) respectively. This means that the rotation of image accompanies easily with the change of its HOG descriptors. Hence, we have to substantially enhance the robustness of HOG descriptors. Otherwise applications of HOG features would be limited to some narrow ranges.



(a) The original block-size image (b) Image (a) rotated clockwise 90°

Fig. 1. The robustness of HOG descriptors with respect to image rotation.



(a) Annular spatial cells (b) Radial gradient transform

Fig. 2. Illustration of rotation-invariant HOG descriptors.

B. Our approach:

Now, the question is how to significantly improve the robustness of traditional HOG descriptors, when the task regions are drastically rotating. There are many existing feature representation methods based on histogram, which are claimed to have presented the robust and invariant descriptor approaches for classification systems. Currently, two of the most popular and representative ones are 2D HOG [6] and HOG 3D [7], which are interesting solutions to the problems above.

Nevertheless, the bottleneck problems also exist in these approaches: 2D HOG descriptor is inspired by Jhuang *et al.*'s approaches [16] that use 2D Gabor-filter responses combined with optical flow. Such dense representations avoid some of the problems discussed above, but cannot solve these problems completely. Moreover, it brings further time complexity because 2D HOG requires a region of interest (ROI) around the task region, which is usually obtained by using either a separate detector or background subtraction followed by blob detection; Inspired by SIFT descriptor [1], HOG 3D constructs a platonic solids system using auxiliary coordinate system to achieve the intention of invariant feature representation. It is an interesting solution yet with high computational time and memory cost. Although they further distribute their task images (faces) over the 2D polar coordinates and make all task images be congruent in order to reduce memory cost, the computational speed is still a bottleneck. Furthermore, HOG 3D have to rely on the integral videos [7], which limits HOG 3D in some restricted application areas. Therefore, these approaches cannot be considered as complete solutions to the above problems.

These histogram based approaches, such as HOG [8], 2D HOG [6] and HOG 3D [7], *etc.*, are most closely related to this paper. Although significant progress has been made in these works, yet invariant representation for features has not been successfully solved. In order to present a more practical and ideal solution, in this paper, we propose a novel feature descriptor on histograms of oriented gradients, *i.e.*, rotation-invariant histograms of oriented gradients (Ri-HOG), which owns an annular spatial cells type blocks (see Fig. 2(a)). This form of blocks is reminiscent of C-HOG [8], but note that essentially, the extraction approaches of C-HOG feature descriptors are the same as the R-HOG's [8], which even limit the describing ability of C-HOG. Therefore, simply changing the rectangle-type blocks into circle-type blocks cannot make HOG be rotation-invariant. Unlike C-HOG, we not only use annular spatial cells to replace rectangular cells, but also compute these cells on a dense radial gradient as feature descriptors to achieve the invariant and robust feature representation. By doing so, the auxiliary coordinate systems or additional detection approaches are not necessary. Instead what we need is that we only focus on the radial gradient transformations of pixels then calculating the gradient magnitude and the orientation of these radial gradients. Hence the time complexity can not be increased but the invariant representation ability of the features can be extremely enhanced.

In this paper, we adopt radial gradient to represent the gradient for HOG descriptors, which is derived from Takacs *et al.*'s rotation-invariant image features [9]. But different from Takacs *et al.*'s approach, we only use the radial gradient to replace the Gaussian gradient function of conventional HOG. We subdivide the local patch into annular spatial cells (see Fig. 2(a)). How to calculate these descriptors is shown in Fig. 2. In Fig. 2(b), $\forall$ a point p in the circle c, the task is to compute the radial gradient magnitude of point p (x, y). Decompose the vector g into its local coordinate system as $(g^T r, g^T t)$, by projecting g into the r and t orientations as shown in Fig. 2(b). Since the component vectors of g in r and t orientations can be quickly obtained by $r = \frac{p-c}{\|p-c\|}$, $t = R_{\frac{\pi}{2}} r$, we can obtain the gradient g easily on the gradient filter. And, R_θ is the rotation matrix by angle θ.

Since Takacs *et al.* focus on image tracking applications, the speed is more important, they use Approximate RGT and ROC curve to compute the feature descriptors [9]. However, by do doing, it will decrease the distinctiveness of feature descriptors for recognition applications. In order to keep the distinctiveness of feature descriptors for recognition application, we do not follow Takacs *et al.*'s way to abandon gradient magnitudes, cells, and blocks representation system. Therefore, essentially, the feature (Ri-HOG) that we adopt here is an improved HOG feature, but the approach proposed by Takacs *et al.* is a very excellent and novel feature representation method for image tracking applications, which cannot be considered as a type of HOG feature. Ri-HOG persists and develops the discriminative representation of conventional HOG features. Meanwhile, it can also significantly enhances the descriptors with respect to rotation-invariant ability. Simply, we use the following four steps to extract the Ri-HOG descriptors:

1. Subdivide the local patch into annular spatial cells as shown

TABLE I
AVERAGE ACCURACY OF ANNULI SPATIAL CELL WITH DIFFERENT BINS ON PASCAL VOC 2007.

Total of bins	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
mAP (%)	5.6	9.8	16.0	21.4	22.3	23.2	25.1	25.9	27.3	28.4	28.5	29.6	32.7	32.6	27.9	27.6

in Fig. 2(a);
2. Calculate the radial gradient $(g^T r, g^T t)$ of each pixel in the cell;
3. Calculate the radial gradient (RG) magnitudes and the orientations of radial gradients using the Eq. 1:

$$\begin{aligned} M_{\mathrm{RG}}(x,y) &= \sqrt{(g^T r)^2 + (g^T t)^2}, \\ \theta(x,y) &= \arctan\frac{g^T t}{g^T r}; \end{aligned} \tag{1}$$

4. Accumulate the gradient magnitude of radial gradient for each pixel over the annular spatial cells into 13 bins, which are separated according to the orientation of radial gradient. In this way, we can extract the feature descriptors from a dense annular spatial bin of these uniformly spaced cells.

Now, we still have a query why this HOG feature is rotation-invariant. See Fig. 2 (b) again, assuming the local patch has been rotated by an angle, $\forall\ \theta$. Point $p \rightarrow$ point p', which generates a new gradient system $R_\theta p = p';\ R_\theta r = r';\ R_\theta t = t';\ R_\theta g = g'$. We can verify the coordinates of the gradient in point p' can be expressed by $(g^T r, g^T t)$:

$$\begin{aligned} (g'^T r', g'^T t') &= ((R_\theta g)^T R_\theta r, (R_\theta g)^T R_\theta t) \\ &= (g^T R_\theta^T R_\theta r, g^T R_\theta^T R_\theta t) \\ &= (g^T r, g^T t). \end{aligned} \tag{2}$$

In such a way, all rotated points in the local patch also can obtain their coordinates of the gradient from the corresponding original points, because all gradients are rotated by the same angle θ, they are one-to-one mapping. Thus, the set of gradients on any given circle or annular spatial bin centered around the patch is invariant.

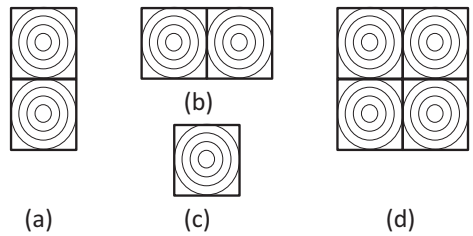


Fig. 3. Some configuration examples for local patches.

C. Blocks Normalization:

Illustration of Fig. 3 shows four main configuration examples for our local feature patches. In every local patch, each one square sub-window is described on a set of descriptors extracted from the Ri-HOG block. The patch size depends on the size of the block, which ranges from 50 $\times$ 50 to 100 $\times$ 100 by sliding the patch over the recognition template by 5 pixels forward. We further allow different aspect ratio for each patch (the ratio of width and height). In this way, it can ensure enough feature-level difference for robust representation according to different images. We adopt $L_2 - Hys$, L_2 normalization followed by clipping [8] as our block normalization approach, because it turned out to work best in practice.

III. EXPERIMENTS

A. Databases and Implementation Details

The experiments are evaluated on PASCAL VOC 2007 dataset [5], which includes 9,963 images of 20 different object classes, containing 5,011 training images and 4,952 testing images. It is the most popular dataset for object detection/recognition and many evaluation experiments of state-of-the-art methods are carried out on it. But note that it is very difficult to improve the results on this dataset. The latest top results are almost the same.

B. Classifiers

Our experiments is implemented in C++ on the RHEL (Red Hat Enterprise Linux) 6.5 OS platform by the PC with Core i7-2600 3.40 GHz CPU and 8 GB RAM.

By default, we use a soft ($C = 0.01$) linear SVMs trained with SVMLight [8] (slightly modified to reduce memory usage for problems with large dense descriptor vectors). Adopting a Gaussian kernel increases 13.0% at the cost of recognition time (18.9 ms pre frame), but it does not provide the increase in performance.

C. Experimental results

The discriminative ability of the proposed method not only depends on its descriptors, but also relates to the number of bins in each annuli spatial cell. Generally, more spatial bins increase distinctiveness, but it will lead to sparse pixels at each bin, which decreases the robustness. Therefore, it is a trade-off and we have to balance them. Table I shows the the mean average precisions (mAP) of the proposed method adopting the annular cell on different bin numbers. We have observed that when the number of bins greater than 13 on PASCAL VOC 2007 dataset, the accuracy was not or seldom improved any more. Therefore, we set the number of bins as 13.

We also carry out the experiments to evaluate well-known existing local feature representation methods: HOG, SURF, and SIFT. The mAP of Ri-HOG, HOG, SURF, SIFT, on PASCAL VOC 2007 dataset are 32.7%, 22.1%, 16.8%, and 24.7% respectively. Therefore, Ri-HOG dominates these existing features on the accuracy. Fig. 4 shows the represented results obtained using Ri-HOG. Comparing them, we can find

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON PASCAL VOC 2007 DATABASE.

	Accuracy of different object-category items (%)																					
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	TV	**mAP**
UCI [10]	28.8	56.2	3.2	14.2	**29.4**	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.1
DPM [11]	26.3	59.4	2.3	10.2	21.2	46.2	52.2	7.9	15.9	17.41	10.9	2.9	53.4	37.6	38.2	4.9	16.6	29.7	38.2	40.8	26.6
LEO [12]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
DSO [13]	32.5	60.1	11.1	16.0	31.0	50.9	**59.0**	**26.1**	21.2	**26.5**	25.4	**16.4**	**61.7**	48.3	42.2	16.1	**28.2**	30.1	44.6	46.3	34.7
CA [14]	34.5	**61.1**	**11.5**	**19.0**	22.2	46.5	58.9	24.7	21.7	25.1	27.1	13.0	59.7	51.6	44.0	**19.2**	24.4	33.1	48.4	**49.7**	**34.8**
AFP [15]	24.1	54.7	1.6	9.8	20.0	42.1	50.1	8.0	13.8	16.7	8.9	2.5	49.4	38.3	36.0	4.2	14.9	24.4	35.8	35.0	24.5
Ours	**35.3**	59.4	5.3	16.8	20.1	**54.6**	52.3	15.2	**22.5**	18.6	**31.0**	16.2	53.5	**52.0**	**47.1**	6.3	19.7	**33.4**	**48.8**	45.8	32.7

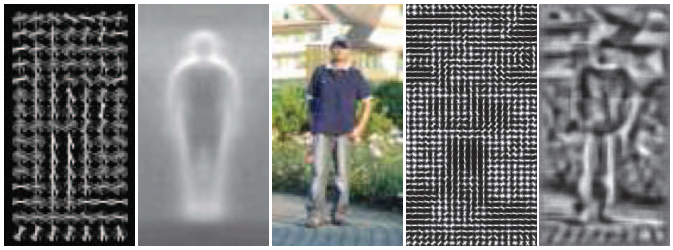


Fig. 4. The detectors comparison on silhouette contours. (a) The original image represented on the traditional HOG [8]; (b) The average gradient image of the traditional HOG descriptors [8]; (c) The original image [8]; (d) The original image represented on the Ri-HOG; (e) The average gradient image of the Ri-HOG descriptors.

that the sample represented by the proposed method is much more discriminative and clear, because the radial gradient can represent the image gradient with all orientations. However, as the conventional ones adopt Gaussian gradient function (in x- and y-), they cannot describe the image gradient with all orientations accurately.

In Table II, experimental results are carried out on comparisons of our approach and six state-of-the-art methods (UCI [10], DPM [11], LEO [12], DSO [13], CA [14], and AFP [15]). Our mAP is 32.7%, including 8-item tasks with the top performance, which are highly competitive to these state-of-the-art methods: 1 item [10], 0 item [11], 0 item [12], 6 items [13], 5 items [14], 0 items [15].

The reason why Ri-HOG is significant process for classification applications is that this invariant features descriptors can make the subject independent classification system more robustness. Because when using subject independent algorithms, there will be noise present in the alignment. Furthermore, almost all current classification frameworks are the typical cases that adopt subject independent algorithms. Hence, invariant features are extremely important for the current classifiers.

IV. CONCLUSIONS

In this paper, we have proposed a novel feature representation method, called rotation-invariant histograms of oriented gradients (Ri-HOG), for the image invariant representation. Furthermore, the discriminative ability and validity of Ri-HOG, also have be experimentally proved in this paper. Therefore, Ri-HOG will be important to those with closely related research interests.

About the future work, we will further study the question about the impact caused by our choice of feature space on recognition.

REFERENCES

[1] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis. (IJCV)*,vol.60, pp.91-110, 2004.
[2] B. Herbert, T. Tinne and V. G. Luc, SURF: Speeded Up Robust Features, in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp.404-417, 2006.
[3] C. Strecha, A.M. Bronstein, M.M. Bronstein and P. Fua, "LDAHash: Improved Matching with Smaller Descriptors", *IEEE Trans. PAMI,* vol.34, pp.66-78, Jan. 2012.
[4] M. Brown, H. Gang and S. Winder, "Discriminative Learning of Local Image Descriptors", in *IEEE Trans. PAMI,* vol.33, pp.43-57, Jan. 2011.
[5] M.Everingham, L. V. Gool, C. K. Williams, J. Winn, , and A. Zisserman,"The Pascal Visual Object Classes (VOC) Challenge" *Int. J. Comput. Vis. (IJCV)*, vol.88, pp.303-338, 2010.
[6] C. Thurau, and V. Hlavac, "Pose primitive based human action recognition in videos or still images", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1-8, Jun. 2008.
[7] A. Klaser, M. Marszae, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients"*Proc. British Machine Vis. Conf. (BMVC)*, pp.275:1-10,2008.
[8] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.886-893, Jun. 2005.
[9] G. Takacs, V. Chandrasekhar, S. S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Fast Computation of Rotation-Invariant Image Features by an Approximate Radial Gradient Transform", *IEEE Trans. Image Proc.*, vol.22 pp.2970-2982, Aug. 2005.
[10] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative Models for Multi-class Object Layout", *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp.229-236, Sept. 2009.
[11] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models", *IEEE Trans. PAMI*, vol.32, pp.1627-1645, Sept. 2010.
[12] L. Zhu, Y. Chen, A. Yuille, W. Freeman, D. McAllester and D. Ramanan , "Latent Hierarchical Structural Learning for Object Detection", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1062-1069, Jun. 2010.
[13] X. Wang, L. Lin, L. Huang, and S. Yan, "Incorporating Structural Alternatives and Sharing into Hierarchy for Multiclass Object Recognition and Detection", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.3334-3341, Jun. 2013.
[14] F. S. Khan, R. M. Anwer, J. Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color Attributes for Object Detection", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.3306-3313, Jun. 2012.
[15] P. Dollar, R. Appel, S. Belongie, and P. Perona , "Fast Feature Pyramids for Object Detection", *IEEE Trans.PAMI*, vol.36, pp.1532-1545, Aug. 2014.
[16] H. Jhuang, T. Serre, L. Wolf and T. Poggio, "A Biologically Inspired System for Action Recognition", *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp.1-8, Oct. 2007.