HOG-Based Descriptors on Rotation Invariant Human Detection

Panachit Kittipanya-ngam and Eng How Lung

Institute for Infocomme Research, 1 Fusionopolis Way,
21-01 Connexis (South Tower), Singapore 138632

Abstract. In the past decade, there have been many proposed techniques on human detection. Dalal and Triggs suggested Histogram of Oriented Gradient (HOG) features combined with a linear SVM to handle the task. Since then, there have been many variations of HOG-based detection introduced. They are, nevertheless, based on an assumption that the human must be in *upright* pose due to the limitation in geometrical variation. HOG-based human detections obviously fails in monitoring human activities in the daily life such as sleeping, lying down, falling, and squatting. This paper focuses on exploring various features based on HOG for rotation invariant human detection. The results show that square-shaped window can cover more poses but will cause a drop in performance. Moreover, some rotation-invariant techniques used in image retrieval outperform other techniques in human classification on *upright* pose and perform very well on various poses. This could help in neglecting the assumption of *upright* pose generally used.

1 Introduction

Because of the demand of smart surveillance system, the research on human detection has gained more attention. Not only is it a fundamental function required in most of surveillance system but also a challenging task in computer vision. In the past decade, there have been many proposed techniques on human detection. Enzweiler and Gavrila [1] review and decompose human detection into three stages: the generation of initial object hypotheses (ROI selection), verification (Classification) and temporal integration (Tracking). They also evaluate the state-of-the-art techniques in human detection: Haar wavelet-based AdaBoost cascade [2], Histogram of Oriented Gradient (HOG) features combined with a linear Suport Vector Machine (SVM) [3], neural network using local receptive fields [4] , and combined hierarchical shape matching and texture-based Neural Network using local receptive fields classification [5]. In this paper, we focus on studying characteristic of features passed to classifiers in the stage of classification.

In 2005, Dalal and Triggs [3] suggested to use HOG features combined with a linear SVM to handle the human body detection. Since then, there have been many variations of HOG-based detection introduced. They are, however, all based on a major assumption that the target human must be in *upright* pose.

R. Koch et al. (Eds.): ACCV 2010 Workshops, Part I, LNCS 6468, pp. 143–152, 2011.

It is mentioned in [3] that HOG descriptor is limited to a certain range of geometrical variation not bigger than the bin size. While the transition variation can be solved by scanning detection windows through whole image and scale variation can be managed by multi-scale methods, rotation variation is still in doubt. Therefore HOG-based body detection is limited to only such applications as detecting human in group photos, detecting and tracking human walking in the scene, and detecting human actions in *upright* pose [6]. This is why the HOG-based human detection fails in the task of monitoring human activities in the daily life such as sleeping, lying down, falling, and squatting.

This paper is focusing on exploring various HOG-based descriptors on rotation invariant human detection. Section 2 briefly explains HOG and discusses why its variations can not be invariant to rotation transformation. A review of rotation invariant features is described in Section 3. Finally, exploratory experiments on various HOG-based features for rotation invariant human detection are illustrated in Section 4 5 and 6 before they are discussed and concluded in section 7.

2 HOG: Histogram of Oriented Gradients

In 2005 [3], Navneet Dalal and Bill Triggs proposed a descriptor representing local object appearance and shape in an image, called Histogram of Oriented Gradient (HOG). The HOG descriptor is described by the distribution of edge directions in the histogram bins. The common implementation begins by dividing the detection window into small a square pixels area, called cells, and for each cell estimating a histogram of gradient directions for those pixels within the cell. The final descriptor is the combination of all histograms in the detection window. In [3], Dalal and Triggs suggest to use 64x128 pixel detection window and 8x8 pixel cell. For detecting human in an image, Support Vector Machine (SVM) is introduced to handle the task of human/non-human classification in each detection window by training SVM with human and non-human images. Since HOG was introduced in CVPR 2005, it has been widely used for detecting human, modified to obtain a better performance and extended to various applications [7,6,8,9]. Though there have been many variations of HOG human detectors proposed, those techniques assume the *upright* position of human because they can not handle the rotation variation. Here we suggest two possible explanations why HOG-based techniques are not rotation invariant: the features change a lot when image is rotated and the shape of detection window does not support the oriented version of object.

In [3], Dalal and Triggs also explain that they chose to define the detection window at 64 x 128 pixels, a rectangle shape because it includes more or less 16 pixels around the person from every side. They have shown that the size of border is important to the performance as it is expected to provide the right amount of context which can help in detection. They tried to reduce the size of pixels around the person from 16 to 8 and obtain the detection window of size 48 x 112 pixels. They have found that the performance of the detector with

48 x 112 detection window is 6% worse than detection window of size 64 x 128 pixels. They also tried to maintain the size of window at 64 x 128 pixels while increasing the person size in it and they have found that it also causes a similar drop of performance. Later, every work on HOG follows this idea of the shape and the size of detection window.

The *upright* rectangle shape of detection window is obviously a reason why HOG can not handle rotation transformation because this shape can not contain other poses of human inside especially rotated version of human such as sleeping, lying down and falling. In this paper, we suggest to use square-shaped detection window as the square window can contain more variations of human. However, we have to be awared that the area of context pixels of square shape will be more than that of *upright* rectangle. We studied how the bigger amount of context information would effect the performance in 5.

3 Review of Rotation Invariant Features

Though not many techniques tackling on orientation in human detection have been proposed, there have been many suggestions on rotation invariant features on other objects especially for the task of image retrieval. Mavandadi *et.al* [10] suggest to construct a rotation invariant features from magnitude and phase of Discrete Fourier Transform of polar-transformed image. Islam *et.al* [11] transform image using curvelet transform, an extended of 2-d ridgelet transform. Then the transformed features were aligned to the main dominant direction to generate the output which is invariant to the rotation of image. Jalil *et.al* [12] align the features by maximise the probability of the vectors obtained from Radon Transform(RT) and Discrete Differential Radon Transform(DDRT). Marimon and Ebrahimi [13] introduce Circular Normalised Euclidean Distance(CNED) to help in aligning the image orientation based on histogram of gradient orientation. Izadinia *et.al* [14] assume the boundary of object is clearly given and then using relative gradient angles and relative displacement vectors to construct a look up table (R-table) in Hough Transform, which is rotation-invariant. Arafat *et.al* [15] study the geometrical transformation invariance of several descriptors in the task of logo classification. Four features compared in this paper are Hu's Invariant Moment[16], Hu's moments of log-polar transformed image, Fourier transformation of log-polar transformed image and Gradient Location-Orientation Histogram (GLOH, a SIFT descriptor on log-polar coordinate). The similarity measure used in this paper is Euclidean Distance. Peng [17] applies Discrete Fourier Transformation (DFT) on HOG to reduce the circular shift in the image and measure the similarity by L1 metric distance equation.

Pinheiro introduces Edge Pixel Orientation Histogram (EPOH) in [18]. This technique divides each image into NxN subimages. On each subimage, HOG is applied to extract the distribution and then feature vector of each subimage is concatenated to construct the final feature. In this technique, the angle considered is within 0 and 180 degree. Therefore, the pixel of edge orientation outside this range will be counted in the bin of opposite angle and the normalised by

the number of edge pixels in the subimage. EPOH is quite similar to HOG but they are different in the number of blocks and the way to construct histogram. Later in [18], they suggest to use Angular Orientation Partition Edge Descriptor (AOP) in image retrieval as it is invariant to rotation and translation. Given a centre point of image, AOP divides the gradient image into N angular sectors of the surrounding circle. In each angular division, the orientation of edge pixel is adjusted by using the angle of the radial axis as the reference to construct local angular orientation. Hence, the radial axis is the line drawn from centre point of the image to the centre point of the sector. Next, histogram is applied to each angular division to extract the local distribution of angular orientation before the feature vector of each angular sector is concatenated to construct 2-D vector $f(n_0, n_a)$ where n_0 is the bin of local angle and n_a is the angle of the radial axis for each sector. The final descriptor will be the absolute value of 1-D Fourier Transform of $f(n_0, n_a)$ relatively to the angular dimension n_a.

In this paper, we studied on five HOG-based techniques as follows:

HOGwoBlk: Histogram of Oriented Gradients without block division
HOGwoBlk-FFT: Amplitude of Fourier Coefficients of HOGwoBlk[17]
HOGwoBlk-FFTp: Phase of Fourier Coefficients of HOGwoBlk[17],
EPOH: Edge Pixel Orientation Histogram[18].
AOP: Angular Orientation Partition Edge Descriptor[18].

Hence, in our implementation, HOGwoBlk does not divide image in blocks and cell as in [3] because EPOH is considered a kind of HOG with divisions. This is for studying the effect of structural information on the performance. Additionally, on *upright* rectangle window, AOP was divided into 2 x 4 block divisions instead of angular divisions. However, the idea of angular sector and the idea of distribution of local orientation were still maintained by assigning the radial axis of each block to be the line drawn from centre point of the image to the centre point of the block. The rest of process remained the same as the original AOP in [18].

4 Various Features on Rotation Invariance

Here we assume that if the feature is rotation invariant, the extracted features of the target image and any of rotated versions should be very similar. In this section, the similarity between the target image and rotated versions of images is studied through various features. Here we selected six square images, 128 x 128 pixels, shown in the top line of figure 1. Five features mentioned in 3 were then extracted and used as reference features. Next, each image was rotated 15 degree counter clockwise and processed to obtain the features. In this section, each image was rotated for 24 times, 15 degree counter clockwise each time, to reach 360 degree. Finally, similarity values between reference images and rotated images on five different features were measured, recorded and plotted against the number of time they were rotated in figure 1. The similarity measurement used in this section is 1D correlation as this measure is similar to the way a linear SVM classifier constructs kernel matrix. In figure 1, the diagrams of original image and

image subtracted foreground are very similar on both upright and lying down images while they are very much different from those of image subtracted background. Clearly, the background information dominate in the feature vectors. It is also noticed that HOGwoBlk-FFT and AOP can highly conserve the similarity over the change of orientation, EPOH is more sensitive than HOGwoBlk-FFT and AOP while HOGwoBlk and HOGwoBlk-FFTp can barely maintain the similarity when rotated. This explains and supports that only histogram of oriented gradients alone is not rotation invariant and why most of work assume *upright* pose but adding some strutural information like EPOH or AOP can improve the performance. Moreover, Figure 1 provides some hints that HOGwoBlk-FFT and AOP would make better features in human detection in the scenario of activities in daily life as they could handle the variation on rotation.

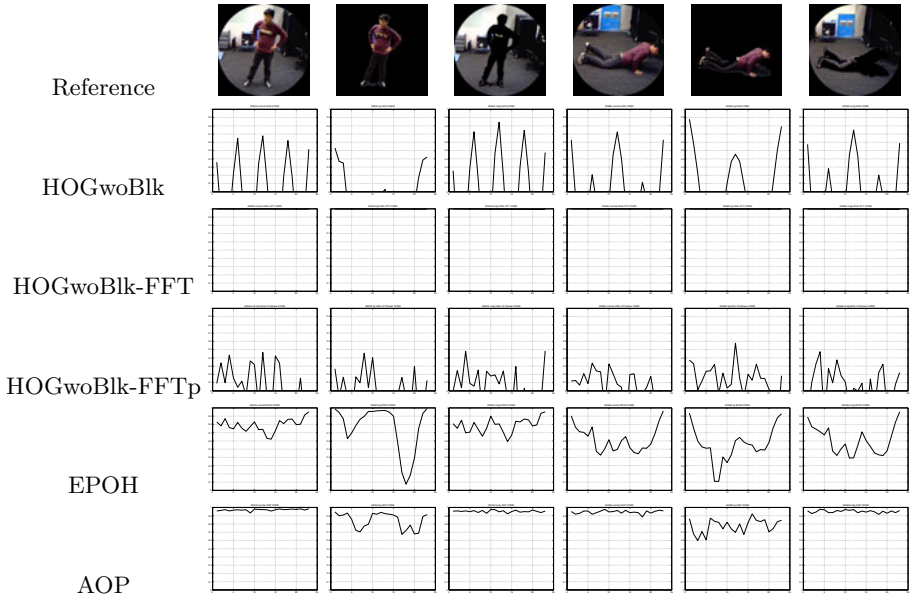
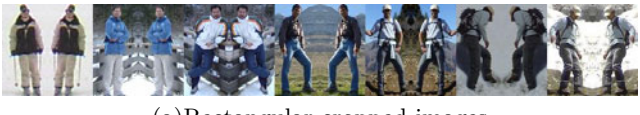


Fig. 1. The correlation between reference image and rotated images, over 24 step of 15 degree, on various HOG-based features

5 Effect of Shape of Detection Window on Presence/Absence of Human Classification

Mentioned previously, the square shape of detection window can cover more poses of human than *upright* rectangle. In this section, the comparison between *upright* rectangle and square shape of detection window is studied on a linear SVM classifier for the task of Presence/Absence of human classification. For each image, the image was cropped and resized into two difference shapes of image, 64x128 rectangle and 128x128 square. For positive images, the cropped area is

based on centroid and boundary where annotations come with the database. For negative images, the windows were randomly cropped and resized. Next, five selected processors mentioned in previous section were applied to extract features from cropped images before the features were used to train and test on classification. Then, the performance of each processors on diffefent shapes of detection window was measured and shown in figure 2 via ROC curve and accuracy of classification in table 1. The database used in this section is called INRIA Person database [3]. The database includes two sets of *upright* photos, for training and for testing and each set is consist of positive images and negative images. Each positive image was flipped around vertical axis for increasing number of possible images. However, some images were discarded in this section because the square window, expanded from *upright* rectangle, covers the area outside the image where there is no pixel info. Therefore the number of images used for training is 2258 images in total, of which 1040 images are positive and 1218 images are negative while the number of images used for testing is 963 images, 510 positive and 453 negative. Some of cropped and resized images used in this section are shown in figure 2. In figure 2(a) and table 1, it is noticed that EPOH and AOP are nearly perfect in the classification when using *upright* rectangle detection window. When using square-shaped detection window, the performance in figure 2(b) and table 1 show that HOGwoBlk-FFT and HOGwoBlk-FFT phase are worse than others. The performance of HOGwoBlk, EPOH and AOP are overall so close to each other but AOP is significantly sensitive as it does not perform well on negative images. The results show that using square-shaped window can cause a drop in overall performance on HOG-based human classification. Probably, the increase of context information in square-shaped window adds more variations in the features and makes the task more difficult to classify for classifiers. Though AOP is slightly better than EPOH when using rectangle-shaped window, EPOH is overall better than AOP when using square-shaped window.



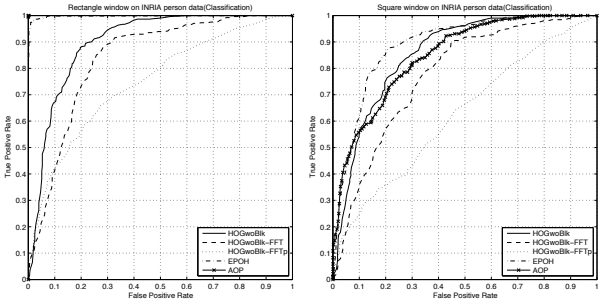
(a)Rectangular-cropped images

(b)Square-cropped images

Fig. 2. Examples of cropped and resized positive images from INRIA Person Database

Table 1. Percentage of correct classification(True Positive(TP) and True Negative(TN)

Type	Rectangle			Square		
	TP	TN	Total	TP	TN	Total
HOGwoBlk	87.84%	80.13%	84.22%	78.04%	76.16%	77.15%
HOGwoBlk-FFT	87.84%	71.74%	79.96%	73.53%	68.65%	71.24%
HOGwoBlk-FFTp	57.25%	79.47%	67.71%	37.45%	77.26%	56.18%
EPOH	98.24%	98.01%	98.13%	78.82%	**84.77%**	**81.62%**
AOP	**100%**	**100%**	**100%**	**82.35%**	68.21%	75.70%



(a)Rectangle window (b)Square window

Fig. 3. ROC of the classification on INRIA person data

6 Rotation Invariant Classification of the Presence of Human

In this section, five focused features will be tested on images of various poses of the activities in daily life, here called CVIU LAB. While INRIA person database was used to train the classifiers as in section 5, the set of various poses used for testing was recorded and annotated by the author, shown in figure 4. In thispart, each image was cropped and resized into square shape of image, 128x128 pixels. Positive images in the test set were sampled from a sequence of human doing a normal daily activities such as stretching arms, falling, lying down and squatting. The cropped area is based on centroid and boundary of foreground obtained from the background subtraction algorithm by Eng [19]. There are 118 images with horizontal flipped versions of them experimented, 236 images in total. For the 220 negative images, the windows were randomly cropped and resized from images without human. After cropped and resized, the feature vector of each selected technique was extracted and pass to pre-trained linear SVM classifiers to decide whether there is a human inside. Then, the performance of classification was measured, shown in figure 5 and discussed.

Figure 5 shows that AOP and EPOH can handle image of various poses when the HOGwoBlk is completely lost. The performance of features can be ranked

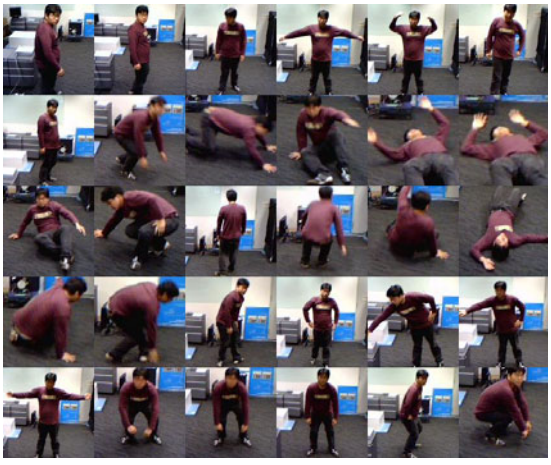
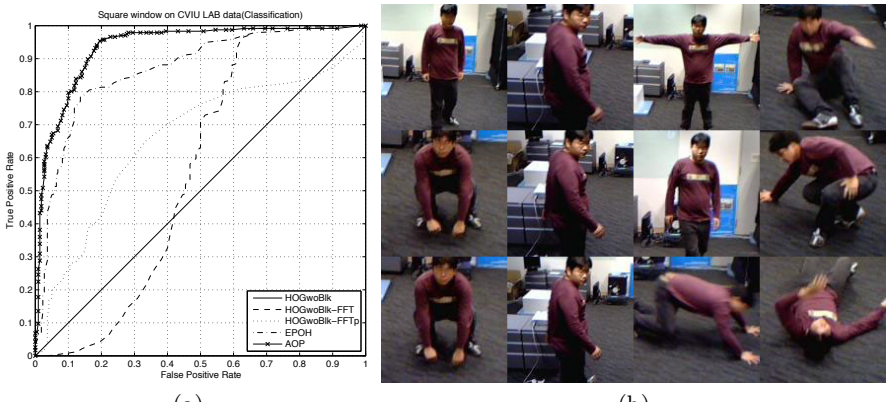


Fig. 4. Examples of cropped and resized images of various poses



(a) (b)

Fig. 5. ROC of the classification on CVIU LAB data

from the best as follows: AOP, EPOH, HOGwoBlk-FFT, HOGwoBlk-FFTp, and HOGwoBlk but HOGwoBlk-FFT is more sensitive in classifying human and non-human. This rank corresponds to the figure 1 showing that AOP, EPOH, and HOG-FFT can maintain similarity over rotation transformation and HOGwoBlk-FFTp and HOGwoBlk change much when image is rotated. Figure 5 displays the images which AOP missed in detecting human inside. It is still in doubt why the AOP can not detect human inside these image while it can handle the other similar images in the data set such as images in figure 4. Hence, all of images in figure 4 are the those AOP could handle.

7 Discussion and Conclusion

Hitogram of orientation of gradients alone can not be used on rotation invariant human detection because of the shape of windows and the significant change of feature over rotation variation. Here we suggest to use square shape of detection window and other kind of features to allow human detection to detect human in various poses. Though it causes a drop in performance.

Applying Fourier Transform relatively to the angular dimension to edge-orientation seems to be about to make edge-orientation features tolerant to rotation change when the similarity is measure by 1-D correlation.

Dividing image into subimages either in angular divisions or blocks can improve overall performance. Probably dividing allows features to include global structural information and the local distribution of edge orientation. This could be a reason why EPOH and AOP outperform other features in human classification.

AOP is the only feature in this study applyinh fouriere transform to reduce the effect of rotation and divided into subimages. This could be the reason why AOP outperform others in the human classification on both *upright* pose and various poses. Its performance convinces that features these characteristics could help neglecting the assumption of *upright* pose generally used in human detection. Though AOP looks nearly perfect to be used for human detection, AOP have higher rate in false positive than EPOH with square-shape window. Presumably, AOP is sensitive to context information in background.

References

1. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 2179–2195 (2009)
2. Viola, P., Jones, M.: Robust real-time face detection. In: Proceedings of the Eighth IEEE International Conference on Computer Vision ICCV 2001, vol. 2, p. 747 (2001)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, vol. 1, pp. 886–893 (2005)
4. Wohler, C., Anlauf, J.: An adaptable time-delay neural-network algorithm for image sequence analysis. IEEE Transactions on Neural Networks 10, 1531–1536 (1999)
5. Gavrila, D.M., Munder, S.: Multi-cue pedestrian detection and tracking from a moving vehicle. International Journal of Computer Vision 73, 41–59 (2007)
6. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp. 1–8 (2008)
7. Lu, W.L., Little, J.: Simultaneous tracking and action recognition using the pca-hog descriptor. In: The 3rd Canadian Conference on Computer and Robot Vision 2006, p. 6 (2006)
8. Li, M., Zhang, Z., Huang, K., Tan, T.: Rapid and robust human detection and tracking based on omega-shape features. In: 16th IEEE International Conference on Image Processing (ICIP 2009), pp. 2545–2548 (2009)

9. Kaaniche, M., Bremond, F.: Tracking hog descriptors for gesture recognition. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance AVSS 2009, pp. 140–145 (2009)
10. Mavandadi, S., Aarabi, P., Plataniotis, K.: Fourier-based rotation invariant image features. In: 16th IEEE International Conference on Image Processing (ICIP 2009), pp. 2041–2044 (2009)
11. Islam, M., Zhang, D., Lu, G.: Rotation invariant curvelet features for texture image retrieval. In: IEEE International Conference on Multimedia and Expo, ICME 2009, pp. 562–565 (2009)
12. Jalil, A., Cheema, T., Manzar, A., Qureshi, I.: Rotation and gray-scale-invariant texture analysis using radon and differential radon transforms based hidden markov models. Image Processing, IET 4, 42–48 (2010)
13. Marimon, D., Ebrahimi, T.: Orientation histogram-based matching for region tracking. In: Eighth International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 2007, p. 8 (2007)
14. Izadinia, H., Sadeghi, F., Ebadzadeh, M.: Fuzzy generalized hough transform invariant to rotation and scale in noisy environment. In: IEEE International Conference on Fuzzy Systems FUZZ-IEEE 2009, pp. 153–158 (2009)
15. Arafat, S., Saleem, M., Hussain, S.: Comparative analysis of invariant schemes for logo classification. In: International Conference on Emerging Technologies ICET 2009, pp. 256–261 (2009)
16. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8, 179–187 (1962)
17. Peng, J., Yu, B., Wang, D.: Images similarity detection based on directional gradient angular histogram. In: Proceedings of 16th International Conference on Pattern Recognition, vol. 1, pp. 147–150 (2002)
18. Pinheiro, A.: Image descriptors based on the edge orientation. In: 4th International Workshop on Semantic Media Adaptation and Personalization SMAP 2009, pp. 73–78 (2009)
19. Eng, H.L., et al.: Novel region-based modeling for human detection within highly dynamic aquatic environment. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004, vol. 2, pp. II–390–II–397 (2004)