PEDESTRIAN DETECTION USING MIXED PARTIAL DERIVATIVE BASED HISTOGRAM OF ORIENTED GRADIENTS

Ali Mahmoud, Ahmed El-Barkouky, James Graham and Aly Farag

ECE Department
University of Louisville,
Louisville, KY 40292, USA
{ali.mahmoud, ahmed.elbarkouky, james.graham, aly.farag}@louisville.edu

ABSTRACT

Recently, several approaches for pedestrian detection have been investigated using discriminatively trained part based models with which Histogram of Oriented Gradients (HOG) showed to be a robust feature. In this paper, we propose a new feature based on HOG to be used with the discriminatively trained part framework for pedestrian detection. Our method is based on computing the image mixed partial derivatives to be used to redefine the gradients of some pixels and to reweigh the vote at all pixels with respect to the original HOG. Our approach was tested on the PASCAL2007 and INRIA person dataset and showed to have an outstanding performance.

Index Terms— Pedestrian Detection, Mixed Partial Derivative, Histogram of Oriented Gradients

1. INTRODUCTION

The computer vision literature has been rich in algorithms for human detection which has many valuable applications in robotics, surveillance, and pedestrian detection for driving assistance [1, 2]. For instance, improving pedestrian detection techniques for driving assistance can significantly reduce accidents and save lives [3].

Recently, several approaches for pedestrian detection have been investigated using discriminatively trained part based models [4]. In these approaches, an object detection system is achieved using mixtures of multi-scale deformable part models that are discriminatively trained using Support Vector Machines (SVM) requiring only the knowledge of the box bounding the object in the image. Felzenszwalb et al. [4] investigated using the Histogram of Oriented Gradients (HOG) features introduced by Dalal and Triggs [5]. The basic idea behind the HOG is that the object local appearance and shape can be described using the distribution of the local intensity gradients or edge directions without the need to precisely know the position of the corresponding gradient or edge [5].

Ren and Ramanan [6] replaced the HOG feature in the framework of [4] by Histograms of Sparse Codes (HSC). In their approach, they formed local histograms by computing sparse codes with dictionaries learned using K-SVD which outperformed the HOG based approach.

Jun et al. [7] introduced the Local Gradient Patterns (LGP) feature and the Binary Histogram of Oriented Gradients (BHOG) feature. The LGP makes the local intensity variations along the edge components robust by assigning one if the gradient of a neighboring pixel is greater than the average of the gradients of the eight neighborhood pixels and zero otherwise. For the BHOG, one is assigned if the value of the histogram bin is greater than the average value of the total histogram bins and zero otherwise. They used their approach for human detection using the INRIA [5] and the Caltech [1, 2] human databases.

Prioletti et al. [8] presented a two-stage pedestrian detection system. Their system used a Haar cascade classifier to extract candidate which are then fed into a part-based histogram of oriented gradient classifier [4] to lower the number of false positives.

Of concern in this paper, we propose a new feature based on HOG to be used with the discriminatively trained part based models human detection framework [4] which plays an important role in recent pedestrian detection systems [8]. The proposed feature is based on computing the image mixed partial derivatives to be used to redefine the gradients of some pixels and to reweigh the vote at all pixels with respect to the original HOG. The mixed partial derivative can be interpreted as the rate of change of the slope in the x direction as we move into the y direction or vice versa which carries information different than that of the derivative in a single direction as x or y.

2. HISTOGRAM OF ORIENTED GRADIENTS

The original HOG was presented by Dalal and Triggs [5]. Throughout this paper, we will focus only on their rectangular version of the HOG. The rectangular HOG was based on locally calculating normalized histograms of image intensity gradient orientations. The local object shape and appearance can be well characterized using these histograms without need to precisely know the position of the image intensity gradients.

To extract the HOG from a window of an RGB colored image, the gradients in the x and y directions are calculated for each color channel separately with the best performance achieved when using 1-D $[-1, 0, 1]$ masks for gradient computation [5]. The resulting gradients are used to compute the gradient orientation and magnitude for each color separately at all pixels and then selecting the values at the color channel corresponding to the highest gradient magnitude at a certain pixel to be fed to the next stage of the HOG computation. Each image window is divided into smaller square patches called cells [5]. An orientation histogram is computed for each cell as follows.

Consider a certain cell; each pixel contributes a weighted vote for a gradient orientation histogram depending on the orientation of the corresponding gradient. The orientation bins of the histogram are either evenly spaced over 0º-180º (for unsigned gradient HOG) or 0º-360º (for signed gradient HOG). The vote at a certain pixel is a function of the gradient magnitude where the best results are achieved when taking the vote equal to the gradient magnitude at the pixel under consideration [5]. After that, a normalization process is done [5] by grouping cells into larger overlapping blocks where each block is normalized separately. This normalization is done in order to compensate for local illumination variations. According to [5], a good performance for human detection can be achieved using nine histogram orientation bins spaced over 0º-180º and normalized over four neighboring blocks leading to a 36-dimensional feature vector.

Felzenszwalb et al. [4] implemented the HOG features by combining unsigned gradient HOG (9 bins spaced over 0º-180º) and signed gradient HOG (18 bins spaced over 0º-360º) in a single feature vector. This leads to a 31-dimensional vector, 27 of them corresponds to these orientations and the remaining 4 captures the overall gradient energy in the four neighboring blocks. Throughout this work, we will use the HOG implemented by [4] as a reference and will build on it.

3. MIXED PARTIAL DERIVATIVE BASED HOG (PROPOSED WORK)

Our method is based on computing the mixed partial derivative of the image window under consideration where the resulting values are used to redefine the x and y components of the gradient of some pixels and to reweigh the vote at all pixels with respect to the original HOG [4, 5].

Let $I_c(x, y)$ denote the image intensity at pixel (x, y) for color channel c, where $c \in \{R, G, B\}$ taking into account that R, G, and B refers to the red, green and blue color channels respectively. Although there are many ways to compute the gradients, using 1-D $[-1, 0, 1]$ masks works best as suggested in [5]. Thus for pixel (x, y) of color channel c as shown in Fig. 1, the gradient in the x direction $I_{c_x}(x, y)$ and the gradient in the y direction $I_{c_y}(x, y)$ are calculated as follows [5]:

$I_c(x-1, y-1) = N_{11}$	$I_c(x, y-1) = N_{12}$	$I_c(x+1, y-1) = N_{13}$
$I_c(x-1, y) = N_{21}$	$I_c(x, y) = N_{22}$	$I_c(x+1, y) = N_{23}$
$I_c(x-1, y+1) = N_{31}$	$I_c(x, y+1) = N_{32}$	$I_c(x+1, y+1) = N_{33}$

Figure 1: 8-connected neighborhood pixels for $I_c(x, y)$.

$$I_{c_x}(x, y) = N_{23} - N_{21}. \quad (1)$$

$$I_{c_y}(x, y) = N_{32} - N_{12}. \quad (2)$$

The mixed partial derivative for pixel (x, y) of color channel c $I_{xy}(x, y)$ is computed as:

$$I_{c_{xy}}(x, y) = (N_{33} - N_{31}) - (N_{13} - N_{11}). \quad (3)$$

Although using 1-D [-1,0,1] masks for computing gradients in the x, and y directions works best as mentioned in [5], we found that recomputing the gradient values at some pixels using pixel values of the neighbors added some improvement. The pixels at which the x and y gradients to be recomputed are based on their original x and y gradients and their mixed partial derivatives. It was clear that some pixels have low values for the absolute values of both the x and y components of the gradient although they have a high value of the mixed partial derivative. As an extreme example, at a given pixel (x,y), $I_{c_x}(x, y) = I_{c_y}(x, y) = 0$ while $I_{c_{xy}}(x, y) \neq 0$. This gives unspecificity when calculating $I_{c_y}(x, y)/I_{c_x}(x, y)$ to compute the corresponding gradient orientation.

To remove this unspecificity at this pixel, the x and y components of the gradient are recomputed using the intensity values at the neighbor pixels. We found that this recomputation improves the performance not only for the extreme case when $I_{c_x}(x, y) = I_{c_y}(x, y) = 0$, but also when $|I_{c_x}(x, y)| < T_x$ and $|I_{c_y}(x, y)| < T_y$ while $|I_{c_{xy}}(x, y)| > T_{xy}$, where T_x, T_y and T_{xy} are positive thresholding values. The recomputation is done as follows:

$$I_{c_x}(x, y) = \begin{cases} \frac{N_{33} + N_{13}}{2} - \frac{N_{31} + N_{11}}{2} & \text{when } |I_{c_{xy}}(x, y)| > T_{xy}, \\ & |I_{c_y}(x, y)| < T_y \text{ and } |I_{c_x}(x, y)| < T_x \\ N_{23} - N_{21} & \text{otherwise} \end{cases} \quad (4)$$

$$I_{c_y}(x,y) = \begin{cases} \frac{N_{33}+N_{31}}{2} - \frac{N_{13}+N_{11}}{2} & \text{when } \left|I_{c_{xy}}(x,y)\right| > T_{xy}, \\ & \left|I_{c_y}(x,y)\right| < T_y \text{ and } \left|I_{c_x}(x,y)\right| < T_x \\ N_{32} - N_{12} & \text{otherwise} \end{cases} \quad (5)$$

For each pixel (x,y), the gradient magnitude $r_c(x,y)$ and the gradient orientation $\theta_c(x,y)$ are calculated for color channel c as follows:

$$r_c(x,y) = \sqrt{I_{c_x}^2(x,y) + I_{c_y}^2(x,y)}. \quad (6)$$

$$\theta_c(x,y) = \text{atan}(\frac{I_{c_y}}{I_{c_x}}), \quad (7)$$

where atan() is the inverse tangent function. The resulting gradient magnitude and the corresponding mixed partial derivative are used to calculate a voting function $v_c(x,y)$ for channel c at pixel (x,y) as follows:

$$v_c(x,y) = \begin{cases} r_c(x,y)\left|I_{c_{xy}}(x,y)\right|^n & \text{when } \left|I_{c_{xy}}(x,y)\right| > T_{xy}, \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where n is a non-negative constant. For each pixel (x,y), c corresponding to the highest $v_c(x,y)$ is picked to be used for an orientation histogram computation. In the following discussion, $v(x,y)$ will denote the value of the largest $v_c(x,y)$ at pixel (x,y) resulting in having a vote array that has number of columns and rows equal to that of the original image window. After computing $v(x,y)$, the vote array is divided into smaller square patches called cells [5]. The size of a cell will be denoted by w_{cell}.

The next step is to compute the histogram for each cell. Consider a certain cell; each pixel contributes a weighted vote $v(x,y)$ for a gradient orientation histogram depending on the orientation of the corresponding gradient. The orientation bins of the histogram are once evenly spaced over 0°-180° using 9 bins and once evenly spaced over 0°-360° using 18 bins and the resulting values are combined in a manner similar to that presented in [4], where the feature vector of each cell is normalized in neighboring square blocks of four cells [5]. In practice, this leads to a 31-dimensional final feature vector.

4. EXPERIMENTAL RESULTS

The performance of the proposed method is evaluated on the PASCAL2007 [9] and INRIA [5] publically available person datasets using the discriminatively trained part based framework of [4] with two components. The cell size w_{cell} is taken as 8×8 pixels. These pedestrian datasets contain images and annotation bounding boxes which represent the ground truth for a detection system [4]. When testing a detector, the input to the system is some images and the output is a set of bounding boxes with corresponding scores [4]. These score can be thresholded at different values to plot the precision-recall curve where

$$Precision = \frac{tp}{tp+fp}. \quad (9)$$

$$Recall = \frac{tp}{tp+fn}, \quad (10)$$

where tp is the number of true positives (correct detections), fp is the number of false positives (unexpected detections), and fn is the number of false negatives (missing detections). At a certain threshold, the precision represents the fraction of the bounding boxes that are correct detections while the recall represents the fraction of the pedestrians in the image that are detected correctly.

Throughout this work, a PASCAL measure has been used to determine the detection rates [9]. If there is an overlap between a detected bounding box and the ground truth bounding box and this overlap is more than 50%, we will consider this a correct detection, otherwise it is a false positive detection. For a certain ground truth bounding box, if there are more than one overlapping detection bounding boxes, only one of them is counted.

We tested several values of the thresholding constants in Eq. (4) and Eq. (5) and selected some values that gave good results. The selected values are $T_x = 0.9$, $T_y = 0.9$, and $T_{xy} = 0.9$. Moreover n in Eq. (8) was chosen as $n = 0.125$.

Fig. 2 shows the precision-recall curves comparing the proposed method with [4] using the PASCAL2007 person dataset. Fig. 3 shows the precision-recall curves comparing the proposed method with [4] using the INRIA person dataset. For both datasets, our results outperform [4].

The proposed method was implemented on our robotic vehicle the ATRV shown in Fig. 4. The ATRV is a Linux based robotic vehicle platform equipped with visible cameras for vision and two 2.4 GHz Quad core computers for processing. Fig. 5 shows sample detections obtained by our system.

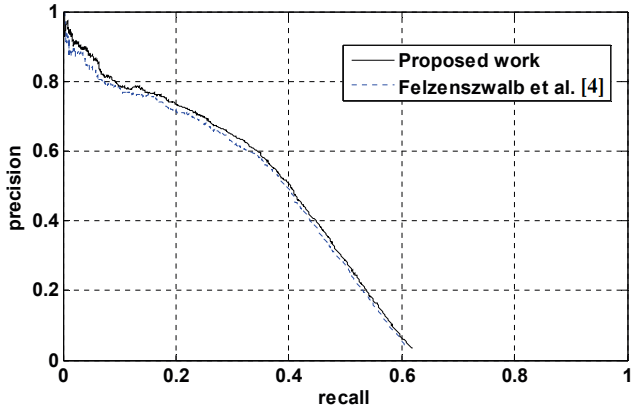


Figure 2: Precision-recall curves for PASCAL2007 person dataset.

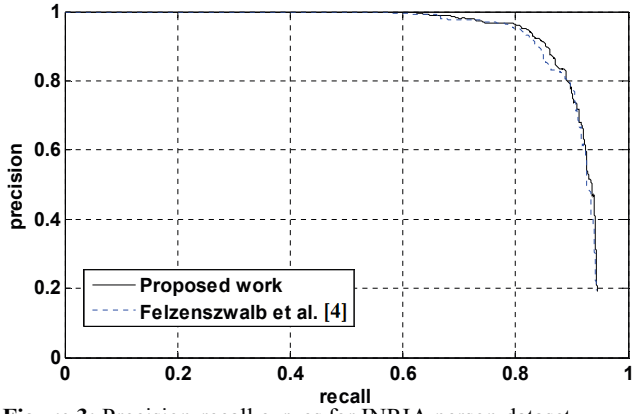


Figure 3: Precision-recall curves for INRIA person dataset.

Figure 4: Our robotic vehicle ATRV.

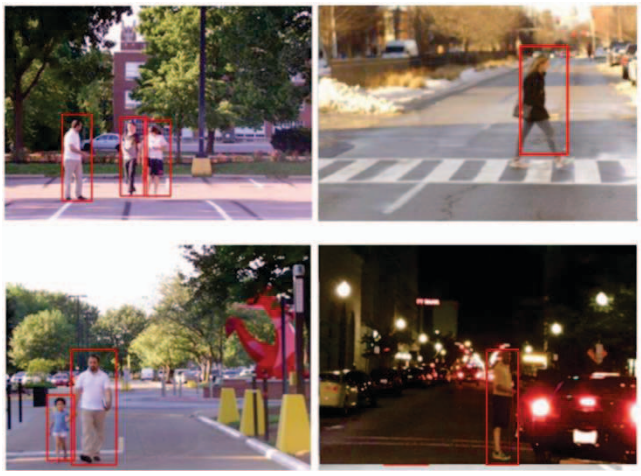
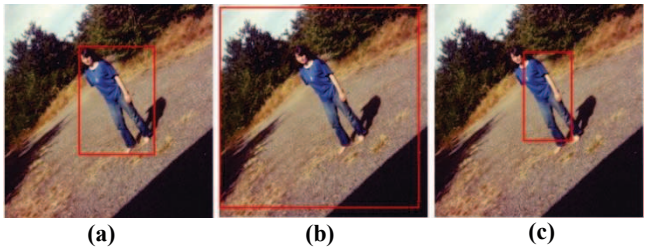


Figure 5: Sample detections of the proposed method.



(a) **(b)** **(c)**

Figure 6: Inclined image from the PASCAL2007 person dataset: (a) Bounding box showing the ground truth. (b) Bounding box showing the detection obtained by Felzenszwalb et al. [4]. (c) Bounding box showing the detection obtained by the proposed method which has an overlap with the ground truth more than 50%.

Fig. 6 shows an inclined image form the PASCAL2007 person dataset. The proposed method succeeded to have an overlap with the ground truth more than 50% and thus was considered as a correct detection while focusing on the subject more than [4]. Since inclined images could be captured by a vehicle moving on an inclined surface, succeeding in detecting pedestrians in these inclined images can be valuable to alert the driver and save lives.

5. CONCLUSION

In this paper, we proposed a new feature to be used with the discriminatively trained part framework for pedestrian detection. Our method makes use of the mixed partial derivatives of the image intensity. We tested the system on two publically available pedestrian datasets and the performance showed to be promising.

6. REFERENCES

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 304-311, June 2009.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, no.4, pp. 743-761, April 2012.

[3] http://safety.fhwa.dot.gov/ped_bike/

[4] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no.9, pp.1627-1645, Sept. 2010.

[5] N. Dalal, and B.Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, pp. 886-893, June 2005.

[6] Xiaofeng Ren, and D. Ramanan, "Histograms of Sparse Codes for Object Detection," IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 3246-3253, June 2013.

[7] Bongjin Jun, Inho Choi, and Daijin Kim, "Local Transform Features and Hybridization for Accurate Face and Human Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.6, pp. 1423-1436, June 2013.

[8] A. Prioletti, A. Mogelmose, P. Grisleri, M.M. Trivedi, A. Broggi, and T.B. Moeslund, "Part-Based Pedestrian Detection and Feature-Based Tracking for Driver Assistance: Real-Time, Robust Algorithms, and Evaluation," IEEE Transactions on Intelligent Transportation Systems , vol.14, no.3, pp. 1346-1359, Sept. 2013.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.