ABSTRACT

Pedestrian detection has been seen huge progress in recent years, much thanks to the Histograms of Oriented Gradients (HOG) features. However, this method (HOG and SVM) has a large number of false detections. To conquer the problem, we provide an affirmative answer by proposing and investigating a salience representation for pedestrian detection, Histograms-Of-Salience (HOS). We extracted saliency map learned from data by using Histogram Based Contrast, and aggregate salient value and oriented gradients to form local HOS. We intentionally keep true to the sliding window framework and only change the underlying features. By learning and using local HOS feature that are much more expressive than HOG, we demonstrate large improvements on the public INRIA dataset.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval.

General Terms
Algorithms, Experimentation

Keywords
Pedestrian detection, Histograms-Of-Salience (HOS), Histogram Based Contrast

1. INTRODUCTION

Object detection is a fundamental problem in computer vision with wide applications such as surveillance, image retrieval, robotics and intelligent vehicles. Since pedestrian detection is one of the most important topics in object detection, it has attracted much attention in recent years.

Meanwhile it is one of the most challenging vision tasks, due to the great variety of appearance and shapes of human figures, highly cluttered backgrounds, and often low quality image sources. Many pedestrian detection methods [1, 2, 3, 4, 5] have been proposed to detect the objects in still image by sliding windows. There has been huge progress in object detection, much thanks to the celebrated Histograms of Oriented Gradients (HOG) features proposed by Dalal-Triggs [6]. However, these methods with HOG features have a large number of false detections (as shown in Fig.1(a)). They take full use of oriented and gradients features from local regions, but ignore the salient information. There are so many evidences that salience features can avoid these false detections (as shown Fig.1(b)). We can see that false detections always come from non-salient regions. Therefore, we proposed HOS feature, based on the salient value of the region, applied in the pedestrian detection.

We can see from previous work [7] that there are two im-
portant components in the pedestrian detection. First, features capture the most discriminative information of pedestrians. Second, a classifier decides whether a candidate window shall be detected as enclosing a pedestrian and SVM (Support vector machine) is often used. The connection between features and classifier components is usually achieved using manual parameter configuration. The HOG feature is individually designed with its parameters manually tuned given the linear SVM classifier [6]. Then HOG feature becomes fixed when people design new classifiers [8]. A few HOG feature parameters are tuned in [9] and fixed, and then different part models are learned in [10]. By fixing HOG features and deformable models, occlusion handling models are learned in [11], using the part-detection scores as input. However, a majority of detectors surveyed in [2] remain high false rate.

As shown in Fig. 2, the motivation of this paper is to establish automatic interaction in learning oriented gradients and salient value of local region. Well-learned features help to locate parts, meanwhile, well-located parts help to learn more discriminative features for different parts. Our method formulates the learning of two key components into a unified framework. The main contribution of this paper is the local representation which can be effectively learned for pedestrian detection and the learned rich features outperform HOG by large margin as demonstrated on the INRIA dataset.

2. HISTOGRAMS OF SALIENCE

The first step of calculation feature detectors in image preprocessing is to ensure normalized color and gamma values. Then, we will develop Histograms-Of-Salience (HOS), which resembles HOG but is based on salient regions that represent each local patch using gradient, orientation and salient value.

The per-pixel salient value is extracted by histogram based contrast [12]. Once per-patch gradient, orientation and salient value are computed, we aggregate them into 'histograms' on regular cells and use them to replace HOG. We divide the image into small connected regions called cells (8×8), and compile a histogram of gradient directions with salient value for the pixels within the cell. The combination of these histograms then represents the descriptor (HOS). For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block (16×16), and then using this value to normalize all cells within the block.

Salient Value. According to the Histogram Based Contrast method which define saliency values for image pixels using color statistics of the input image. The pixels with the same color value have the same saliency value under this definition, since the measure is oblivious to spatial relations. Therefore, it is so easy to get saliency value for each pixel as,

$$ Y(I_{(x,y)}) = \sum \left( d(I_{(x,y)}, I_{(x,y)^*}) \right), $$

where \(d_I\) included in the image \(I\) in color space; \(d(I_{(x,y)}, I_{(x,y)^*})\) is the color distance metric between pixels \(I_{(x,y)}\) and \(I_{(x,y)^*}\) in the LAB space. To be more precise, the new saliency value is the average of the pixel saliency value and eight adjacent pixels \(D\), as \(S(I_{(x,y)})\).

$$ S(I_{(x,y)}) = \frac{1}{9} \sum_{D} Y(I_{(x,y)}), $$

where \(D\) is patch of 3×3, so \(D\) is written as follow,

$$ D = \begin{pmatrix} I_{(x-1,y-1)} & I_{(x-1,y)} & I_{(x-1,y+1)} \\ I_{(x,y-1)} & I_{(x,y)} & I_{(x,y+1)} \\ I_{(x+1,y-1)} & I_{(x+1,y)} & I_{(x+1,y+1)} \end{pmatrix} $$

Gradient Computation. This step is to simply apply the 1-D centered, point discrete derivative mask in one or

![Figure 2: The framework of our pedestrian detection system.](Image)

![Figure 3: Histogram generation of HOG and HOS features. Green histograms represents the portion of the pedestrian area; Blue histograms represents the portion of the background area. HOS histogram feature can distinguish between pedestrians and background.](Image)
both of the horizontal and vertical directions. Specifically, this method requires filtering the color or intensity data of the image with the following filter kernels: $[-1, 0, 1]$ and $[-1, 1, 0]^T$. We can get the gradient value of the pixel $H(x, y)$ as follows.

$$G_x(x, y) = H(x + 1, y) - H(x - 1, y), \quad (4)$$

$$G_y(x, y) = H(x, y + 1) - H(x, y - 1). \quad (5)$$

The magnitude and orientation of gradient can be written as,

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2}, \quad (6)$$

$$\theta(x, y) = \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right). \quad (7)$$

So, the new gradient with salient value can be written as follow,

$$G_s(x, y) = G(x, y)S(x, y), \quad (8)$$

We can see from Fig. 3 that pedestrian in our method is more prominent in histogram than HOG without salient information.

**Orientation Binning.** The main step of calculation involves creating the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the new gradient computation. The cells themselves can either be rectangular or radial in shape, and the histogram channels are evenly spread over 0 to 180 degrees. We use 9 histogram channels in pedestrian detection experiments. As for the vote weight, pixel contribution can either be the gradient magnitude itself, or some function of the magnitude; in actual tests the gradient magnitude itself generally produces the best results. Last, we make the block normalization.

**SVM classifier.** The final step in pedestrian recognition which uses Histogram of Silence (HOS) descriptors is to feed the descriptors into some recognition system based on supervised learning. The SVM classifier is a binary classifier which looks for an optimal hyperplane as a decision function. Once trained on images containing some particular object, the SVM classifier can make decisions regarding the presence of an object, such as a human being, in additional test images. In the Dalal and Triggs human recognition tests, they used the freely available SVMLight software package in conjunction with their HOS descriptors to find pedestrian figures in test images.

### 3. EXPERIMENTS

In this section, the proposed approach is validated by comparing with several pedestrian detection methods (based on HOG) on INRIA dataset. The dataset contains 1774 pedestrian positive examples and 1671 negative images without people. The pedestrian annotations were scaled into a fixed size of $64 \times 128$, which include a margin of 16 pixels around the pedestrians. The dataset was divided into two, where 1,000 pedestrian annotations and 1,000 person-free images were selected as the training set, and 774 pedestrian annotations and 671 person-free images were selected as the test set. For each cascade level, the SVM classifier was trained using all the positive examples and $N_n = 10,000$ negative examples generated by bootstrapping. Detection on the INRIA pedestrian dataset is challenging since it includes subjects with a wide range of variations in pose, clothing, illumination, background, and partial occlusions.

**Figure 4: INRIA per-window results.**

Fig. 4 shows that the performance of our method is comparable to the HOG approaches. We compare ours experiments results with HOG on INRIA dataset. The x-axis corresponds to false positives per window (FPPW), the y-axis corresponds to the miss rate, and we plot the detection error tradeoff curves on a log-log scale for INRIA dataset. We consider both the kernel and linear SVM method of HOG.

**Figure 5: INRIA per-image results.**

Fig. 5 shows that the performance of our method is comparable to the HOG approaches. We compare ours experiments results with HOG on INRIA dataset. The x-axis corresponds to false positives per image (FPPI), the y-axis corresponds to the miss rate, and we plot the detection error tradeoff curves on a log-log scale for INRIA dataset. We consider both the kernel and linear SVM method of HOG.

### 4. CONCLUSIONS

In this work we demonstrated that salience based features, learned from image, can replace and outperform the HOG features for general pedestrian detection. The detection problem is long thought to be a challenging case for feature
learning, with millions of windows to consider. Through effective learning and training, we successfully showed how to build and use Histograms-of-Silence features (HOS) in the spirit of HOG, which are capable of representing rich structures beyond gradients and lead to large improvements on virtually all classes on the INRIA dataset. The main contribution of this paper is the local representation which can be effectively learned for pedestrian detection and the learned rich features outperform HOG by large margin as demonstrated on the INRIA dataset. Our work is the first to clearly demonstrate the advantages of feature learning for general pedestrian detection, which come at a reasonable computational cost. Our studies show that large structures in image patches, when captured in an image of complex scene, generally improve pedestrian detection, calling for future work on designing and learning even richer features.

5. ACKNOWLEDGMENTS

The research was supported by the major national science and technology special projects (2010ZX03004-003-03), the National Basic Research Program of China (973 Program) (2009CB320906), National Nature Science Foundation of China (61231015), the National Natural Science Foundation of China (60970160, 61070080, 61172173, 61303114), China Postdoctoral Science Foundation funded project (2013M530350), the National Science Foundation of Hubei Province of China(2010CD05103), the Fundamental Research Funds for the Central Universities (2042014k0250,2042014k0286,2042014k0212), the science and technology Program of Wuhan (201271031366).

6. REFERENCES