15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Crowd Pedestrian Detection Using Expectation Maximization with Weighted Local Features

Shih-Shinh Huang and Chun-Yuan Chen
Dept. of Computer and Communication Engineering,
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
poww@nkfust.edu.tw

Abstract

This study proposes a method for crowd pedestrian detection based on monocular vision using expectation maximization (EM) with weighted local features. The proposed method mainly consists of two stages: training and detection stages. During training stage, the proposed method firstly constructs a model for describing the pedestrian appearance based on a set of salient features. During detection stage, an algorithm called expectation maximization (EM) is applied to group the extracted corners to several pedestrians based on the constructed codebook through performing E-step and M-step iteratively. The use of EM algorithm makes the proposed method be capable of detecting partially occluded pedestrians, especially in crowded scenes. In the experiment, a well-known dataset called CAVIAR is used to validate the effectiveness of the proposed method.

1 Introduction

As wide deployment of cameras in public environment, such as airports, parting lots, and mass-transit stations, accurately estimating the number of people and locating each individual from a fixed camera are important issues in the area of video surveillance. The relevant methods proposed in the literature to tackle with these challenges can be generally classified into two categories: detection-based approach and feature grouping approach.

The methods in the detection-based category formulate the detection problem as a binary sliding window classification problem. For the purpose of tolerating some degree of occlusion and shape articulation, those models usually use local patch representation to describe the pedestrian appearance. The mostly cited local patch feature is called histogram of oriented gradients (HOGs) proposed by [10] and the robustness of HOGs is discussed in [8]. Besides HOGs, other widely used ones include shapelet [9], edgelet [12], LBP (local binary pattern) [6] and EOH (edge orientation histogram) [3]. Although the methods based on local patch features may achieve accurate detection, they have some limitations that most of them are time-consuming, are only suitable for the scenes with slight occlusion, and require the images with high resolution.

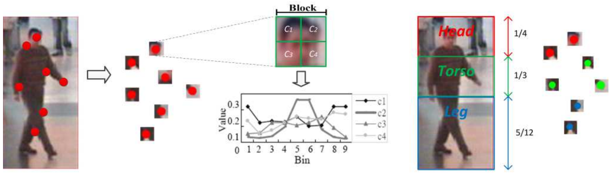
The other category for crowd pedestrian detection is referred to as feature grouping which starts from detecting a large set of salient features followed by partitioning these into a set of individuals using explicit grouping algorithm or model-based segmentation algorithm. The salient features widely used in this category are corner-like feature points [2, 5]. The representation of the detected point is generally its position [5] or gray-encoded vector [11] of the patch centered at that point. One way for feature grouping is through voting mechanism based on implicit shape model (ISM) [11] which specifies where the feature might appear with respect to the object center. However, grouping features using ISM may be failed to detect pedestrians in the scene with seriously occlusion. This is because the vote resulting from the pedestrian with occlusion is insufficient to generate an initial hypothesis. Be available in crowded scenes, expectation maximization (EM) algorithm capable of solving missing data problem is an alternative and widely used way for feature grouping [5, 7]. The EM algorithm groups features by iteratively assigning each feature to cluster(s) (E-Step) and updating the cluster parameters (M-Step). Motivated by the above discussion, the proposed method belongs to the feature grouping category and EM algorithm is applied for the crowd pedestrian detection.

The remainder of this paper is organized as follows. Section 2 introduces how to model the pedestrian appearance by using HOGs as well as spatial property. Section 3 elaborates the detail to group extracted corners into several individuals through EM algorithm. Section 4 analyses the performance of the proposed method using the popular CAVIAR dataset. Finally, Section 5 concludes this paper with some discussion.

2 Part Patch Codebook Construction

Since the pedestrians in crowded scenes are generally occluded of each other, only partial pedestrian appearance is visible in the image. Therefore, the key to successfully detect pedestrians in crowded scenes is the effective modelling of the local pedestrian appearance. In this work, the appearance of local patch centered at the salient point as well as the part label of salient point are modelled as a codeword. Then, a set of representative codewords is collected to model the pedestrian and is referred to as a codebook.

For each corner point cp_i extracted from Harris corner detector [13], a local patch centered at cp_i is denoted as a block B_i and described by HOGs [10]. The resulting 36-D HOGs feature vector is denoted as $\mathcal{V}_i = \{v_k^{(i)}\}_{k=1}^{36}$. In addition to HOGs descriptor, the part label indicating where the corner locates is also encoded in order to impose spatial property. The part label l_i assigned to cp_i is one of three body parts: head (H), torso (T), or leg (L) depending which region the cp_i locates on. The head, torso, and leg regions are defined as regions with respect to $\frac{1}{4}$, $\frac{1}{3}$, and $\frac{5}{12}$ height of the manually annotated pedestrian clip as shown in



(a) HOGs-Based Patch Descriptor (b) Part Region Definition

Figure 1. Corner Description: (a) HOGs feature for representing the patch centered at the corner. (b) The definition of three regions head, torso, and leg, respectively.

Figure 1 (b). In short, each corner point cp_i is represented by $cp_i = \{\mathcal{V}_i, l_i\}$, where $\mathcal{V}_i = \{v_k^{(i)}\}_{k=1}^{36}$ is the 36-D HOGs descriptor for representing the patch centered at cp_i and $l_i \in \{H, T, L\}$ encodes the spatial property of cp_i.

For a set of corners $\{\mathcal{V}_i, l_i\}_{i=1}^{N_{cp}}$, where N_{cp} is the total number of corners extracted from all training images, each is firstly considered as a single-element cluster $C_i = \{cp_i\}, i = 1, 2, .., N_{cp}$. Then, the two clusters with the highest similarity are merged into a new cluster until their similarity is less than a defined threshold Th_C which is set as 0.5. The similarity between two clusters C_m and C_n is defined:

$$F_{sim}(C_m, C_n) = \frac{\sum_{cp_i \in C_m} \sum_{cp_j \in C_n} f_{sim}(cp_i, cp_j)}{|C_m| \times |C_n|} \tag{1}$$

where $|C_m|$ denotes the cardinality of C_m and $f_{sim}()$ measures the similarity between two cluster elements (corner points). The definition of $f_{sim}()$ is through the use of normalized gray-scale correlation (NGC) [5].

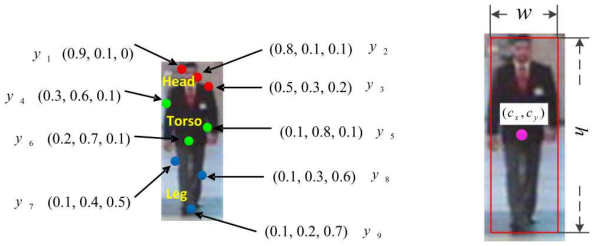
Let $C_m = \{cp_k^{(m)}\}_{k=1}^{|C_m|}$ be a cluster obtained from aforementioned grouping algorithm. The cluster C_m is then represented by a codeword $\mathcal{W}_m = \{\bar{\mathcal{V}}_m, \gamma_m^{(H)}, \gamma_m^{(T)}, \gamma_m^{(L)}\}$. The $\bar{\mathcal{V}}_m$ is the mean vector of HOGs descriptors of all corner points in C_m and is taken as the appearance representative of C_m. $\{\gamma_m^{(l)} : l \in H, T, L\}$ are referred to as the part probabilities of C_m that encode possibility of body parts where the corner points in C_m are from. The definition of $\{\gamma_m^{(l)}\}$ is the proportion of the corner points in C_m with the part label equal to l, that is,

$$\gamma_m^{(l)} = \frac{\#\{cp_k^{(m)} | l_k^{(m)} = l\}}{|C_m|}, l \in \{H, T, L\} \tag{2}$$

Consequently, the codebook $\mathcal{B}$ is the collection of the codewords $\mathcal{B} = \{\mathcal{W}_m\}_{m=1}^{|\mathcal{B}|}$, where $|\mathcal{B}|$ is the number of codewords in $\mathcal{B}$ and is automatically determined by the proposed grouping algorithm.

3 EM-Based Pedestrian Detection

This section will shed light on how to detect pedestrians in crowd using the proposed codebook $\mathcal{B}$ which models pedestrian appearance as well as spatial property.



(a) Part Probabilities Assignment (b) Pedestrian Hypothesis

Figure 2. Notation Definition: (a) Weighted Corners: $y = \{cp, \tilde{\gamma}^{(H)}, \tilde{\gamma}^{(T)}, \tilde{\gamma}^{(L)}\}$ (b) Parameters of a pedestrian hypothesis:$\theta = (c_x, c_y, w, h)$

3.1 Problem Formulation

For every extracted corner point cp, it is then described by the aforementioned HOGs descriptor which is denoted as $\mathcal{V}$ and used to find a matched codeword in the constructed codebook $\mathcal{B}$. A codeword $\tilde{\mathcal{W}}$ which has a match to cp is determined through $NGC(.)$ measure and is defined as:

$$\tilde{\mathcal{W}} = \arg \max_{\mathcal{W}_m \in \mathcal{B}} NGC(\mathcal{V}, \mathcal{V}_m) \tag{3}$$

Hence, the cp is assigned three part probabilities as those of the codeword and is referred to as weighted corner $y = \{cp, \tilde{\gamma}^{(H)}, \tilde{\gamma}^{(T)}, \tilde{\gamma}^{(L)}\}$. The resulting set of weighted corners of the observed image is denoted as $Y = \{y_i\}_{i=1}^{|Y|}$ and used for further pedestrian detection. An example of the nine weighted corners extracted from an example clip and their associated three part probabilities is shown in Figure 2(a).

A pedestrian hypothesis used in this work is a rectangular bounding box which is parameterized as $\theta = (c_x, c_y, w, h)$, where (c_x, c_y) is the coordinate of the center point and w and h are, respectively, the width and height. Accordingly, the estimating the parameters $\Theta = \{\theta_i\}_{i=1}^{|\Theta|}$ of $|\Theta|$ pedestrian hypotheses based on the observation Y leads to the detection of pedestrians. Under probabilistic formulation, the estimation of Θ given Y can be achieved by maximizing the log-likelihood probability $\log \Pr(Y|\Theta)$, that is,

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} \log \Pr(Y|\Theta) \tag{4}$$

Since the pedestrians in the crowded scenes may occlude of each other, it is generally hard to directly formulate $\Pr(Y|\Theta)$. To solve this, a latent variable $Z = \{z_i\}_{i=1}^{|Y|}$ is introduced to model the relation between Y and Θ, where $z_i \in \{0, .., |\Theta|\}$ is a discrete random variable that defines which pedestrian hypothesis the weighted corner y_i from. The use of Z leads $\Pr(Y|\Theta)$ to $\int_Z \Pr(Y, Z|\Theta)$ and a two-step algorithm called EM [4] can be applied to estimate Θ in an iteratively manner.

3.2 Estimation of Initial Pedestrian Hypotheses

Since head is the salient and robust feature [7] in pedestrian detection, especially, in crowded scenes due to its low variance in appearance and high visibility in

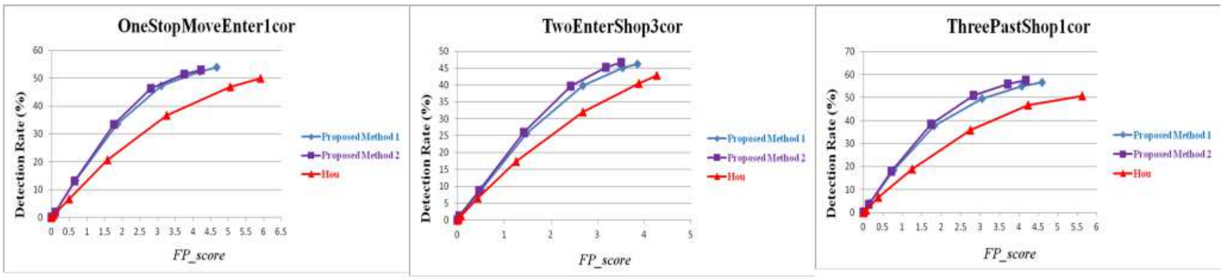


Figure 3. The ROC evaluation for the three videos

different views, the first way firstly identifies the corners locating on the head body part which are called head corners. In this work, the weighted corners with the head probability $\gamma^{(H)} >= 0.5$ are directly considered as head corners. Then, each head corner is used to generate a pedestrian hypothesis by taking it as the top-middle point. For the purpose of dealing with the head invisibility in cases of occlusion and pedestrians far from camera, the second way is to use the corners locating on the torso body part. Similarly, the weighted corners with $\gamma^{(T)} \geq 0.5$ are considered as torso corners. Each generated pedestrian hypotheses takes a torso corner as its center.

3.3 Detection Algorithm

Based on the initial pedestrian models $\Theta^{(0)}$, the E-Step as well as M-Step is applied to iteratively maximize the log-likelihood probability $\Pr(Y|\theta)$. In E-Step, let $R(\theta_j)$ be the rectangular region determined by θ_j and it can be divided into three sub-regions, head region $R^{(H)}(\theta_j)$, torso region $R^{(T)}(\theta_j)$, and leg region $R^{(L)}(\theta_j)$. The belonging relation of a specific weighted corner y_i to these three regions defines the assignment possibility that y_i comes from θ_j as:

$$a_{i,j} = \begin{cases} \tilde{\gamma}_i^{(l)} & \text{if } y_i \in R^{(l)}(\theta_j), l = H, T, or L \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Accordingly, the normalization of a_{ij} leads to the definition of the term $\Pr(z_i = j|y_i, \Theta^{(m)})$ which can be expressed as:

$$\Pr(z_i = j|y_i, \Theta^{(m)}) = \frac{a_{i,j}}{\sum_{j=1}^{|\Theta|} a_{i,j}} \tag{6}$$

For notation simplicity, let $\alpha_{i,j}$ be the short for $\Pr(z_i = j|y_i, \Theta^{(m)})$.

In M-Step, the update of pedestrian hypotheses Θ is to adjust every pedestrian hypothesis θ_j so that those weighted corners assigned to θ_j are well covered in $R(\theta_j)$. For a pedestrian hypothesis θ_j, the centers of its three sub-regions with respect to head, torso, and leg, $\{(c_x^{(l)}, c_y^{(l)}) : l = H, T, L\}$, are respectively calculated using weighted average strategy as follows.

$$(c_x^{(l)}, c_y^{(l)}) = \frac{\sum\limits_{y_i \in R^{(l)}(\theta_j)} \alpha_{i,j} \times cp_i}{\sum\limits_{y_i \in R^{(l)}(\theta_j)} \alpha_{i,j}} \quad l = \{H, T, or L\} \tag{7}$$

The update of center point of θ_j is expressed as:

$$(c_x, c_y) = \sum_{l=H,T,L} (\beta^{(l)} \times (c_x^{(l)}, c_y^{(l)})) \tag{8}$$

where $\alpha^{(l)}$ denotes the weight of the center point at body part l and is defined as:

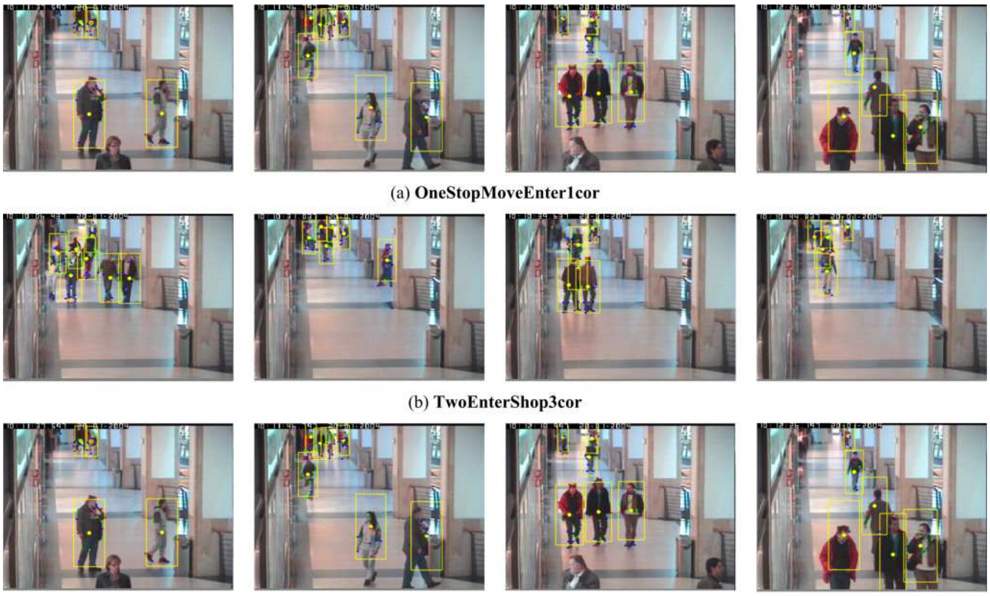
$$\beta^{(l)} = \frac{\sum\limits_{y_i \in R^{(l)}(\theta_j)} \alpha_{i,j}}{\sum\limits_{y_i \in R(\theta_j)} \alpha_{i,j}} \quad l = \{H, T, or L\} \tag{9}$$

4 Experiment

In this section, the 15 videos from a well-known CAVIAR dataset [1] are used to validate the proposed method for crowd pedestrian detection. Among 15 videos, 12 are used for training and the remaining 3 videos are for testing. Of all training images in the training videos, there are 725 pedestrian clips without occlusion. The 2160 corner points in these training clips are then extracted and a codebook consisting of 173 codewords is constructed by grouping the corners with similar HOGs descriptors.

In addition, the method proposed by Hou [5] is also implemented for comparison. In Hou's approach, the detection of individuals in crowded scenes is through the grouping of corner points using EM algorithm without imposing the spatial property of corners. Furthermore, a variant of the proposed method which is referred to as "Proposed Method 1" is used for comparison as well. It only takes the maximal part probability for EM clustering for the purpose of analysing the effect of encoding spatial property as three part probabilities. The originally proposed method is thus referred to as "Proposed Method 2" in later discussion. Figure 4 (a), (b), and(c) show the detection result of some examples of the three testing videos, respectively. Although the pedestrians in the crowded scenes are occluded of each other, in low resolution, or with various poses, the "Proposed Method 2" can successfully identify individuals in the image.

In addition to subjective evaluation, the ROC (Receiver Operating Characteristics) curve which illustrates the relation of detection rate and false positive rate is used for objective evaluation. The three ROC curves of the aforementioned methods for the three testing videos are shown in Figure 3. Obviously, the



(a) OneStopMoveEnter1cor

(b) TwoEnterShop3cor

(c) ThreePastShop1cor

Figure 4. The detection results of the three videos by the "Proposed Method 2"

two methods encoding spatial property of corners outperform the Hou's approach. Furthermore, the encoding of spatial property in a finer manner makes the "Proposed Method 2" obviously superior to the "Proposed Method 1", in cases of the pedestrians with low resolution and various poses.

5 Conclusion

This study proposes a feature grouping algorithm based on a model for detecting pedestrians in crowded scenes. The pedestrian model used here is a codebook whose entry represents the local appearance as well as spatial property of the salient corners. The appearance is described by well-known HOGs; spatial property is encoded as the three part probabilities. The combination of these two features facilitates the generation of initial pedestrian hypotheses and makes EM algorithm achieve better performance in cases of occlusion and low resolution. The experimental results in CAVIAR dataset exhibit that the detection accuracy of the proposed method is about 5% to 10% higher than Hou's one. But, describing the corner using HOGs only has about 57% matching accuracy in average. In the near further, extra cues should be introduced to combine with HOGs for improving the representative ability of corners.

References

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/, 2010.

[2] C. Y. Jeong, S. Choi, and S. W. Han, "A Method for Counting Moving and Stationary People by Interest Point Classification," *IEEE Intl. Conf. on Image Processing* (2013) 4545–4548.

[3] Y. Ma, L. Deng, X. Chen, and N. Guo, "Integrating Orientation Cue with EOH-OLBP-Based Multilevel Features for Human Detection," *IEEE Trans. on Circuits and Systems for Video Technology* **23** (2013) 1755–1766.

[4] Y. Chen and M. R. Gupta, "EM demystified: An Expectation-Maximization Tutorial," *Technical Report, Department of Electrical Engineering, University of Washington* (2010).

[5] Y. L. Hou, and G. K. H. Pang, "People Counting and Human Detection in a Challenging Situation," *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans* **41** (2011) 24–33.

[6] Y. Mu, S. Yan, L. Liu, T. Huang, and B. Zhou, "Discriminative Local Binary Patterns for Human Detection in Personal Album," *IEEE Intl. Conf. on Computer Vision and Pattern Recognition* (2008) 1–8.

[7] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified Crowd Segmentation," *European Conf. on Computer Vision* (2008) 691–704.

[8] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Performance Evaluation of Local Features in Human Classification and Detection," *IET Computer Vision* **2** (2008) 236–246.

[9] Sabzmeydani and G. Mori, "Detecting Pedestrians by Learning Shapelet Features," *IEEE Intl. Conf. on Computer Vision and Pattern Recognition* **1** (2007) 1–8.

[10] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Intl. Conf. on Computer Vision and Pattern Recognition* **1** (2005) 886–893.

[11] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes," *IEEE Intl. Conf. on Computer Vision and Pattern Recognition* **1** (2005) 875–885.

[12] B. Wu and R. Nevatia, "Detectiong of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors," *IEEE Intl. Conf. on Computer Vision* (2005) 90–97.

[13] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Alvey Vision Conference* (1988) 147–151.