2009

Human detection based on weighted template matching

Duc Thanh Nguyen

University of Wollongong, dtn156@uow.edu.au

Philip Ogunbona

University of Wollongong, philipo@uow.edu.au

Wanqing Li

University of Wollongong, wanqing@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Nguyen, Duc Thanh; Ogunbona, Philip; and Li, Wanqing: Human detection based on weighted template matching 2009, 634-637.

<https://ro.uow.edu.au/infopapers/2129>

Human detection based on weighted template matching

Abstract

This paper proposes a new two-stage human detection method involving matching and verification. A Bayesian framework is developed to verify the matching score obtained from a weighted distance measure. Performance evaluation indicates that the proposed method is able to utilize the flexible matching scheme and produce superior true positive, true negative and low misclassification rates.

Keywords

detection, human, template, matching, weighted

Disciplines

Physical Sciences and Mathematics

Publication Details

Nguyen, D., Ogunbona, P. & Li, W. (2009). Human detection based on weighted template matching. IEEE International Conference on Multimedia and Expo (pp. 634-637). USA: IEEE.

HUMAN DETECTION BASED ON WEIGHTED TEMPLATE MATCHING

Nguyen Duc Thanh, Philip Ogunbona, and Wanqing Li

Advanced Multimedia Research Lab, ICT Research Institute
School of Computer Science and Software Engineering
University of Wollongong, Australia

ABSTRACT

This paper proposes a new two-stage human detection method involving matching and verification. A Bayesian framework is developed to verify the matching score obtained from a weighted distance measure. Performance evaluation indicates that the proposed method is able to utilize the flexible matching scheme and produce superior true positive, true negative and low misclassification rates.

Index Terms— Human detection, template matching

1. INTRODUCTION

Human detection from images and videos is a crucial step in human motion analysis and activity recognition. The challenge of the task arises from the numerous variations that human postures can assume and the complexity of the surrounding environment (e.g. cluttered background, crowded scene, etc.). In this paper, the problem of human detection is defined as follows: Given an image, identify all humans and delineate the extent of each by a compact bounding box. In this paper, human postures are supposed to be upper right standing.

A number of approaches employing temporal, appearance and depth information [1] have been proposed in the literature in which shape-based approach holds several advantages. First, shape constitutes a good descriptor for humans in images and videos. Second, shape plays an important role in discriminating human form from other types of objects. It is a robust descriptor when there is no information about human appearance such as colour or texture.

Generally, shape description can be explicitly or implicitly incorporated into human detection algorithms. For example, some authors [2, 3, 4, 5] have explicitly employed full body templates while others [6] used body part templates. This approach has the advantage of allowing for the variations of human postures but requires manual labelling. The implicit use of shape [7, 8, 9, 10, 11], entails learning models based on human/non-human training patterns. Shape features are thus determined automatically without the need for hand labelling. However, this approach limits the possible variations of human poses to what is available in the training data set.

The approach presented in this paper is motivated by the works in [2, 3], and seeks to extend them through the use of prior information and weighted distance computation. In Section 2, we briefly present some previous related works in the literature. The proposed method of human detection is described in Section 3 and experimental results based on USC and MIT pedestrian data sets along with some comparative analysis are presented in Section 4. We discuss the results and draw conclusions in Section 5.

2. RELATED WORK

Explicit use of human shapes relies on manually labelled 2-dimensional contour representation. Gavrilu et al. [2] cluster human templates into a hierarchical structure where the similarity between two templates is defined by the Chamfer distance. For each sliding detection window, the candidate template is selected by traversing the tree from root to leaf in depth-first search fashion. This work has been extended in [3], where a probabilistic method is proposed to compute thresholds used to reject unmatched regions.

Also based on template matching, Lin et al. [6] applies the method in [2] to detect individual body parts including head-torso, upper legs, and lower legs. The full human body is then constructed in a hierarchical structure of detected body parts.

Some methods combine both local and global approaches [4, 5]. The general idea can be summarized as follows. First, codebooks corresponding to local shapes are defined. The relationship between local and global shapes is learned through training. Codebooks are further used to vote for global shapes and detection results are obtained by post processing of maximum votes.

Implicit incorporation of shape in human detection algorithms proceed by learning models from training data. For example, Mohan et al. [7] use Haar wavelets trained by SVM to describe body parts while Wu et al. [8] introduce a so-called "edgelet" feature selected by a real and nested cascade Adaboost algorithm.

In the work of Dalal et al. [9] a histogram based feature named histograms of oriented gradients (HOG) is introduced. HOGs of overlapping areas are concatenated into a vector and

Fig. 1. First row: some templates. Second row (from left to right): input image, edge image, distance image, and the best matched template.

used to train a linear SVM. Extensions of the HOG based shape description are found in [10, 11]. For example, in [10], "Integral Image" is employed to speed up the computation of HOG. Moreover, blocks with variable sizes and cascade Adaboost are used. In [11], edge orientation histogram (EOH) is proposed, in which each vector is a scalar number. A meta-stage is added to cascade Adaboost to exploit the inter-stage information.

3. PROPOSED METHOD

The main idea is inspired by the work of Gavrila et al. [2, 3]. However, we extend this method by the use of prior information and weighted distance which will be described in detail in the next section. The method includes two steps: matching and verification.

3.1. Matching

Given a set of templates describing a number of human postures (see Fig. 1) we find the best matching description for the image within a detection window. The best matching template is that having the shortest distance to the image within the detection window.

Let I_W be the image determined by a detection window W , the best matched template T^* is obtained as,

$$T^* = \arg \min_T D(T, I_W) = \arg \min_T \sum_{t \in T} w_t d_{I_W}(t) \quad (1)$$

where $d_{I_W}(t)$ denotes the distance between pixel (feature) t in T and the closet pixel in I_W (Fig. 1); w_t is the weight assigned to the pixel (feature) t . d_{I_W} can be obtained by computing the Distance Transform.

In [2, 3], w_t is replaced by a constant $1/|T|$, ($|T|$ is the cardinality of T) and thus $D(T, I_W)$ becomes the average distance. However, the average distance does not describe shape locally and will not provide the flexibility to capture the variation of human shapes. In the proposed method the weight, w_t , assigned to each pixel (feature) $t \in T$, is obtained from

Fig. 2. Some examples from: USC dataset (1st row), MIT dataset (2nd row), and negative images (3rd row).

the sigmoid function as,

$$w_t = 1 / \left\{ 1 + \exp \left[- \frac{D_{\bar{S}(T)}(t) - D_{S(T)}(t)}{D_{S(T)}(t) + D_{\bar{S}(T)}(t)} \right] \right\} \quad (2)$$

where

$$D_X(t) = \frac{1}{|X|} \sum_{x \in X} d_x(t)$$

X represents $\bar{S}(T)$ or $S(T)$.

In (2), $\bar{S}(T)$ is the set of negative images while $S(T)$ is the set of positive images corresponding to T . $S(T)$ can be found in two ways. First, it can be obtained manually or alternatively, it can be obtained as the set of positive training images whose the distances to T (using the average distance as in [2, 3]) are smaller than a predefined threshold. Therefore, $D_{\bar{S}(T)}(t)$ indicates the average distance from a feature t to the closest image features of all negative images and $D_{S(T)}(t)$ represents the average distance from t to its nearest image features of all images in $S(T)$, i.e. positive images corresponding to T . The values of w_t are in the range (0, 1) and indicate the importance of the corresponding feature t . Furthermore, the value assumed by w_t is dependent on the margin between $D_{\bar{S}(T)}(t)$ and $D_{S(T)}(t)$. A large margin corresponds to a large w_t and provides a good representation of T at t . If $D_{\bar{S}(T)}(t) = D_{S(T)}(t)$, $w_t = 0.5$, i.e. t is an ambiguous feature. Finally, w_t is normalized so that $\sum_{t \in T} w_t = 1$.

3.2. Verification

Once the best matching template, T^* is found, a verification is required to ascribe a degree of confidence on whether I_W contains a human. In [2, 3], the criterion is based on the matching scores, i.e. $D(T^*, I_W)$. In this paper, we determine the confidence according to two criteria including the matching score and prior information. The prior information encodes the credibility we have in the best matching template. For

example, if a template is often found to be matched with negative samples, it is dubious. The verification process can be stated as a conditional probability,

$$P(Human|T, D) \geq \theta \quad (3)$$

where T and D are respectively, the template and its distance to the detected image; and θ is a predefined threshold. The verification is performed using the best matching template (T^*) and the corresponding distance, $D(T^*, I_W)$. For the sake of simplicity of notation in the sequel, we use T instead of T^* , and D instead of $D(T^*, I_W)$. The distance associated with each template T , is a random variable and also statistically independent of the template. Applying Bayes's theorem to the conditional probability, we have,

$$\begin{aligned} P(Human|T, D) &= \frac{P(T, D|Human)P(Human)}{P(T, D)} \\ &= \frac{P(T|Human)P(D|Human)P(Human)}{P(T)P(D)} \\ &= \frac{P(Human|T)P(T)P(Human|D)P(D)P(Human)}{P(Human)P(Human)P(T)P(D)} \\ &= \frac{P(Human|T)P(Human|D)}{P(Human)} \end{aligned}$$

$P(Human|T)$ can be identified as the prior since it indicates the degree of confidence that a given template T represents human posture. To evaluate $P(Human|T, D)$, we assume that:

$$\begin{aligned} P(Human|T) &= \frac{fp(T) + \varepsilon}{fp(T) + fn(T) + 2\varepsilon}, \\ P(Human|D) &= e^{\beta D}, \\ P(Human) &= P(NonHuman) = 1/2 \end{aligned}$$

where $fp(T)$ and $fn(T)$ respectively denote the relative frequency with which T matches positive and negative training images. The constant ε , is a small positive value set to 0.0001 in our experiment to avoid dividing by zero; β is set to 6/255. We can write the conditional probability as,

$$P(Human|T, D) = \frac{2e^{\beta D}[fp(T) + \varepsilon]}{[fp(T) + fn(T) + 2\varepsilon]} \quad (4)$$

4. EXPERIMENTAL RESULTS

4.1. Evaluation

Performance evaluation of the detection method was carried out on two datasets. The first dataset, USC Pedestrian sets (A, B, C) [8, 12], includes 410 positive images and 500 negative images. We used 200 positive and 200 negative images for training (to compute weights (2) and prior information). The second dataset consists of 500 images from the MIT pedestrian test set [7]. However, since this set includes positive

Fig. 3. ROC curves on dataset 1 (top) and dataset 2 (bottom). ROC curves are generated by varying the values of threshold θ in (3).

images only, we employed the same negative images from the first dataset for testing. As illustrated in Fig. 2, compared with the first dataset, MIT test set is relatively simpler, having mainly subjects in frontal/back view. Some experimental results are shown in Fig. 4. Three performance measures namely, true positive, true negative, and misclassification were recorded. Misclassification is defined as,

$$MC = \frac{FalsePos + FalseNeg}{TruePos + TrueNeg + FalsePos + FalseNeg}.$$

These measures are estimated on both datasets and presented in Table. 1.

4.2. Comparison with other works

It is difficult to directly compare our proposed method with the work of Gavrilu et al. [3] since they used different set of templates for training and constructing the template hierarchical tree. In [3], the criterion for verification is based on the matching scores, i.e. comparing $D(T^*, I_W)$ with a threshold. The threshold is determined based on the level of the hierarchical structure and the density of sliding detection window. In our case, the positive images are cropped to fit the detected humans in the template, hence it is not necessary to find ap-

Table 1. True Positive (TP), True Negative (TN), and Misclassification (MC) rate on different datasets.

Dataset	Measure	Gavrila's method	The proposed method
Dataset 1	TP	75.12%	82.68%
	TN	81.60%	85.80%
	MC	21.32%	15.60%
Dataset 2	TP	82.00%	85.60%
	TN	81.60%	85.80%
	MC	18.20%	14.30%

appropriate threshold. Rather, the best threshold is selected to minimize the misclassification rate. By applying this computation, we can achieve a fair comparison. Fig. 3 presents the ROC curves of the two methods with and without using prior and weighted distance. This figure indicates that the combination of prior and weighted distance (the proposed method) outperforms the use of distance only (as in [2, 3]). The improvement gained is also represented by the increasing true positive and true negative rates, i.e. reducing misclassification (Table. 1).

5. CONCLUSIONS

This paper introduces a two-stage human detection method based on weighted template matching and verification, thereby generalizing the work proposed by Gavrila et al. [2, 3]. We introduced prior information and employed Bayesian framework to verify the result of the template matching. We tested and compared the proposed algorithm with the work in [2, 3] on two different datasets. The experimental results show that the proposed method improved the detection performance.

6. REFERENCES

- [1] T. B. Moeslund, A. Hilton, and V. Krger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, 2006.
- [2] D. M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *Proc International Conference on Computer Vision*, 1999, vol. 1, pp. 87–93.
- [3] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408–1421, 2007.
- [4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 878–885.
- [5] E. Seemann, B. Leibe, and B. Schiele, "Multi-aspect detection of articulated objects," in *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1582–1588.

Fig. 4. Some experimental results. Black rectangles represent true detections while grey rectangles and ellipses correspond to false alarms and miss detections respectively.

- [6] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Proc International Conference on Computer Vision*, 2007.
- [7] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, 2001.
- [8] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc International Conference on Computer Vision*, 2005, pp. 90–97.
- [9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.
- [10] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1491–1498.
- [11] Y. T. Chen and C. S. Chen, "A cascade of feed-forward classifiers for fast pedestrian detection," in *Proc Asian Conference on Computer Vision*, 2007, pp. 905–914.
- [12] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc International Conference on Computer Vision*, 2007.