1-1-2009

A novel template matching method for human detection

Duc Thanh Nguyen
University of Wollongong, dtn156@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Philip Ogunbona
University of Wollongong, philipo@uow.edu.au

A novel template matching method for human detection

Abstract

This paper proposes a novel weighted template matching method. It employs a generalized distance transform (GDT) and an orientation map (OM). The GDT allows us to weight the distance transform more on the strong edge points and the OM provides supplementary local orientation information for matching. Based on the matching method, a two-stage human detection method consisting of template matching and Bayesian verification is developed. Experimental results have shown that the proposed method can effectively reduce the false positive and false negative detection rates and perform superiorly in comparison to the conventional Chamfer matching method.

Keywords

novel, human, method, matching, template, detection

Disciplines

Physical Sciences and Mathematics

Publication Details

Nguyen, D., Li, W. & Ogunbona, P. (2009). A novel template matching method for human detection. International Conference on Image Processing (pp. 2549-2552). Cairo, Egypt: IEEE.

A NOVEL TEMPLATE MATCHING METHOD FOR HUMAN DETECTION

Nguyen Duc Thanh, Wanqing Li, and Philip Ogunbona

Advanced Multimedia Research Lab, ICT Research Institute
University of Wollongong, Australia

ABSTRACT

This paper proposes a novel weighted template matching method. It employs a generalized distance transform (GDT) and an orientation map (OM). The GDT allows us to weight the distance transform more on the strong edge points and the OM provides supplementary local orientation information for matching. Based on the matching method, a two-stage human detection method consisting of template matching and Bayesian verification is developed. Experimental results have shown that the proposed method can effectively reduce the false positive and false negative detection rates and perform superiorly in comparison to the conventional Chamfer matching method.

Index Terms— Human detection, template matching, orientation map

1. INTRODUCTION

Detection of humans from an image or video is a crucial step in human motion analysis. It is a challenging task because of the numerous variations of the human postures and the complexity of the surrounding environment. 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 addition, we assume that the humans to be detected are in upper right position and with different postures.

A common approach to solving the problem is to describe human postures using shape descriptors. For instance, a simple shape descriptor is the binary contour representing the human body shape. The shape descriptors can then be explicitly or implicitly incorporated into human detection algorithms. Explicit use of shape information has the advantage of allowing for the variations of human postures, but requires manually labeled shape contours, referred to as *templates* hereafter. For example, Gavrila et al. [1, 2] used a set of templates representing different up-right postures and organized them into a tree structure. For a detection window, its similarity to a template is calculated using the Chamfer distance and a candidate template is selected by traversing the tree from root to leaf using a depth-first strategy. The candidate template is further verified by comparing the similarity with a threshold [2].

Also based on template matching, Lin et al. [3] detected individual body parts including head-torso, upper legs, and lower legs and then combined the detected body parts hierarchically to validate whether the combined shape is a human or not. In [4, 5], codebooks were used to define local shapes. The relationship between local and global shapes was learned through training.

Implicit incorporation of shape in human detection algorithms involves learning models from training data without a need for the binary templates. For example, Wu et al. [6] employed so-called ”edgelet” feature trained by a real and nested cascade Adaboost algorithm to describe body parts. In the work of Dalal et al. [7] histograms of oriented gradients (HOG) were used. HOGs of overlapping blocks were concatenated into a vector to train a linear SVM. Extensions of the HOG are found in [8] in which ”Integral Image” was employed to speed up the computation of the HOG. Moreover, blocks of variable sizes and cascade Adaboost were used.

Motivated by the works of [1, 2], this paper proposes a novel weighted template matching method which employs a generalized distance transform (GDT) and an orientation map (OM). The GDT allows us to weight the distance transform more on the strong edge points during the distance transform and the OM provides supplementary local information for the matching. Based on the proposed matching method, a two-stage human detection method consisting of template matching and Bayesian verification is developed.

The rest of the paper is organized as follows. Section 2 describes the novel weighted template matching method, Section 3 presents the two-stage human detection algorithm based on the proposed template matching. Experimental results along with some comparative analysis are presented in Section 4. Section 5 concludes the paper with remarks.

2. WEIGHTED TEMPLATE MATCHING WITH ORIENTATION MAP

A common problem of conventional template matching based human detection is the high false positive rate in cluttered backgrounds as illustrated in Fig. 1. This is probably due to the interference of the edges in the background with the Chamfer matching on the distance transform (DT) image. To reduce false positive, we propose a novel template matching

Fig. 1. Left: An image with cluttered background. Center: Binary edge image. Right: distance transform image.

method which employs the following three strategies:

- A generalized distance transform (GDT) [9] is adopted so as to weight more on strong edge points in the creation of distance transform image.
- An orientation map (OM) is created simultaneously with the GDT image. Each value of the OM represents the edge direction of the nearest edge pixel.
- The matching score between a test image and a template is computed jointly from the GDT image and OM with a prior weight for each point in the template.

2.1. Generalized Distance Transform (GDT)

Let $\mathcal{G}$ be a regular grid and $\Psi : \mathcal{G} \rightarrow \mathbb{R}$ a function on the grid. According to Felzenszwalb and Huttenlocher [9], the GDT of Ψ can be defined as,

$$D_\Psi(p) = \min_{q \in \mathcal{G}} \{d(p,q) + \Psi(q)\}, \tag{1}$$

where $d(p,q)$ is some measure of the distance between point p and q in the grid. Intuitively, for each point p we find a point q that is close to p, and for which $\Psi(q)$ is small. For conventional distance transform (DT) of an edge image using $L^2 - norm$, $d(p,q)$ is the Euclidean distance between p and q, and $\Psi(q)$ is defined as

$$\Psi(q) = \begin{cases} 0, & \text{if } (q) \in e \\ \infty, & \text{otherwise} \end{cases} \tag{2}$$

where e represents the binary edge pixels.

Notice the conventional DT does not consider the quality of the edge points in e and a cluttered background often contains many weak edge points. In order to reduce the impact of these weak edge points, we define the $\Psi(q)$ as follows such that more trust is placed on the strong edge points.

$$\Psi(q) = \begin{cases} \frac{1}{\sqrt{I_x^2 + I_y^2}}, & \text{if } (q) \in e \\ \infty, & \text{otherwise} \end{cases} \tag{3}$$

where $I_x = \partial I / \partial x$ and $I_y = \partial I / \partial y$ are the horizontal and vertical gradients of the image I at position q.

Using the algorithm proposed by Felzenszwalb and Huttenlocher [9], the GDT can be computed in $O(knm)$ time, where $n \times m$ is the image's size, k ($= 2$ in our case) indicates the number of dimensions.

2.2. Orientation Map

Let q^* be the closest edge point to the pixel p, that is,

$$q^* = \arg\min_{q \in \mathcal{G}} \{d(p,q) + \Psi(q)\}.$$

and the orientation value at p is defined as,

$$O_\Psi(p) = \arctan(I_{x^*} / I_{y^*}). \tag{4}$$

where I_{x^*} and I_{y^*} are the gradients at q^*. In other words, the orientation of edge pixels will be propagated to their nearest non-edge pixels.

The orientation map $O_\Psi(p)$ provides additional information to match a template with a test image. We can see that, $O_\Psi(p)$ and $D_\Psi(p)$ can be calculated simultaneously without increasing complexity.

2.3. Weighted template matching

Given a template T and a test image I, the matching score or dissimilarity is defined as,

$$D(T,I) = \sum_{t \in T} w_T(t) d_{T,I}(t). \tag{5}$$

where $w_T(t)$ is a weight indicating the importance of the point t in T and $d_{T,I}(t)$ is the dissimilarity between I and T at point t. We define $d_{T,I}(t)$ as,

$$d_{T,I}(t) = \sqrt{D_\Psi^2(t) + \sin^2(O_\Psi(t) - o(t))}, \tag{6}$$

where $o(t)$ is the orientation at point t in T and

$$D_\Psi(t) \leftarrow \exp(-1/max\{\varepsilon, \sqrt{D_\Psi(t)}\}),$$

where ε is a small positive number to avoid dividing by zero. Note that if $w_T(t) \in [0,1], \forall t$ and $\sum_{t \in T} w_T(t) = 1$, $0 < D(T,I) < \sqrt{2}$.

Since T is a binary template, $o(t)$ cannot be obtained directly using the gradient image. Instead, it is approximated by the normal of the contour of the human body as shown in Fig. 2.

In [1, 2], $w_T(t)$ is replaced by a constant $1/|T|$, ($|T|$ is the cardinality of T) and thus $D(T,I)$ 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, $w_T(t)$, assigned to each pixel $t \in T$, is obtained from a set of positive and negative training images.

$$w_T(t) = 1 / \left\{ 1 + \exp\left[- \frac{D_{\bar{S}(T)}(t) - D_{S(T)}(t)}{D_{S(T)}(t) + D_{\bar{S}(T)}(t)} \right] \right\} \tag{7}$$

Fig. 2. Left: Gradient vector of a binary point. Right: Human templates with gradient vectors (grey lines).

where

$$D_{X(T)}(t) = \frac{1}{|X(T)|} \sum_{x \in X(T)} d_{T,x}(t)$$

X represents $\bar{S}$ or S and $d_{T,x}(t)$ is defined as in (6).

In (7), $\bar{S}(T)$ is the set of negative images while $S(T)$ is the set of positive images corresponding to T. $D_{\bar{S}(T)}(t)$ indicates the average distance from a point t to the closest image points of all negative images and $D_{S(T)}(t)$ represents the average distance from t to its nearest points of all images in $S(T)$, i.e. positive images corresponding to T. The values of $w_T(t)$ range into $(0, 1)$ and indicate the importance of the corresponding point t. Furthermore, the value assumed by $w_T(t)$ is dependent on the margin between $D_{\bar{S}(T)}(t)$ and $D_{S(T)}(t)$. A large margin corresponds to a large $w_T(t)$ and provides a good representation of T at t. If $D_{\bar{S}(T)}(t) = D_{S(T)}(t)$, $w_T(t) = 0.5$, i.e. t is an ambiguous feature. Finally, $w_T(t)$ is normalized so that $\sum_{t \in T} w_T(t) = 1$.

3. PROPOSED HUMAN DETECTION METHOD

The proposed human detection method includes two steps: matching and verification. Given a set of templates describing a number of human postures we find the best matching description for the image I_W within a detection window W. The best matching template T^* is determined as,

$$T^* = \arg\min_T D(T, I_W), \tag{8}$$

where $D(T, I_W)$ denotes the matching score.

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 [1, 2], the criterion is based on the matching scores, i.e. $D(T^*, I_W)$. In this paper, we determine the confidence according to both 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,

$$P(Human|T, D) \geq \theta \tag{9}$$

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,

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

$P(Human|T)$ 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:

$$\begin{aligned} &P(Human|T) = \frac{fp(T) + \varepsilon}{fp(T) + fn(T) + 2\varepsilon}, \\ &P(Human|D) = \sqrt{2} - D \text{ (since } 0 < D < \sqrt{2}), \\ &P(Human) = P(NonHuman) = 1/2 \end{aligned}$$

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. We can write the conditional probability as,

$$P(Human|T, D) = \frac{2(\sqrt{2} - D)(fp(T) + \varepsilon)}{fp(T) + fn(T) + 2\varepsilon} \tag{10}$$

4. EXPERIMENTAL RESULTS

In our experiments, we used 46 templates obtained by manually labeling the contours of human body in various postures. A few examples are shown in Fig. 2. The proposed method was evaluated using two popular datasets: USC pedestrian sets (A, C) [6, 10] and INRIA pedestrian test set [7]. The USC dataset has 305 images with 545 humans at different views including frontal/rear and profile. We cropped 200 positive and 200 negative windows for training (to compute weights (Eq.7) and prior information). By scanning these 305 images at different scales (from 0.6 to 1.3), we produced $31,637$ positive and $1,454,322$ negative samples. True positive is determined by comparing the detection window with true detection given in the ground truth using the criteria proposed in [4]. These criteria include the relative distance between bounding box centers with respect to the size of the ground truth box, the cover and overlapping between the detection window and the ground truth. For the INRIA dataset, we used 404 positive images. The same evaluation procedure was employed. The ROCs achieved by the proposed method on the

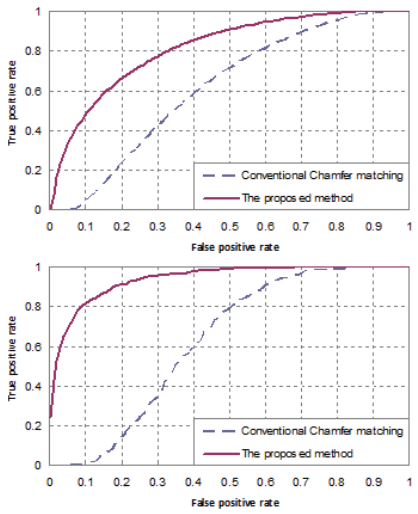

Fig. 3. ROC curves on USC (top) and INRIA dataset (bottom) are generated by varying the values of θ in (Eq. 9). Notice that we do not merge overlapping detection windows.

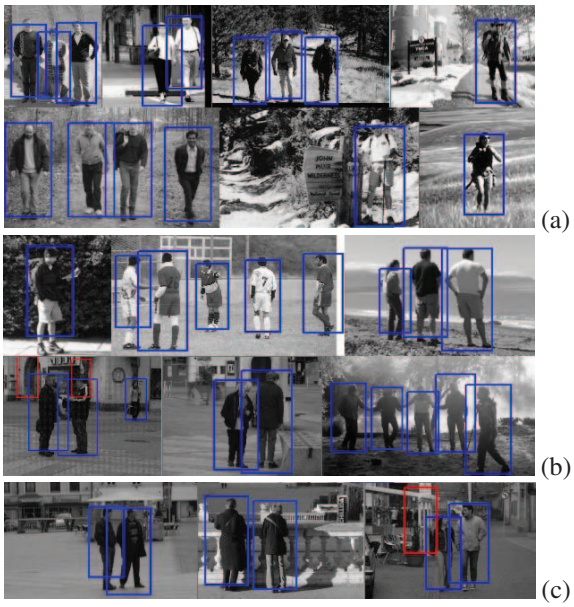

Fig. 4. Some experimental results on USC-A (a), USC-C (b), and INRIA dataset (c). Blue rectangles represent true detections while red rectangles indicate false positives.

two datasets are shown in Fig.3. Some detection results are shown in Fig. 4.

It is difficult to directly compare the proposed method with the work of Gavrila et al. [2] since they used different set of templates. In [2], 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 a detection window. In this experiment, we compare the proposed method and the conventional Chamfer matching used in [1, 2]. Fig. 3 shows the ROCs of these two methods, which clearly demonstrates that the proposed method had much better performance. Since the criterion used to merge overlapping detection windows is not reported in [1, 2], we evaluate the performance for every detection window without merging.

5. CONCLUSIONS

This paper introduces a novel template matching method using GDT and OM. The matching algorithm is then applied to human detection in a two-stage framework. The proposed method generalizes the Chamfer matching method used in [1, 2] in which we introduced the prior information and employed a Bayesian framework to verify the result of the template matching. We tested and compared the proposed algorithm with the conventional Chamfer matching on two different datasets. The experimental results show that the proposed method significantly improved the detection performance.

6. REFERENCES

[1] D. M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *ICCV*, 1999, vol. 1, pp. 87–93.

[2] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *PAMI*, vol. 29, no. 8, pp. 1408–1421, 2007.

[3] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *ICCV*, 2007.

[4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR*, 2005, vol. 1, pp. 878–885.

[5] E. Seemann, B. Leibe, and B. Schiele, "Multi-aspect detection of articulated objects," in *CVPR*, 2006, vol. 2, pp. 1582–1588.

[6] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *ICCV*, 2005, pp. 90–97.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 1, pp. 886–893.

[8] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *CVPR*, 2006, vol. 2, pp. 1491–1498.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," Tech. Rep., Cornell Computing and Information Science, http://www.cs.cornell.edu/~dph/papers/dt.pdf, 2004.

[10] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *ICCV*, 2007.