Robust People Detection by Fusion of Evidence from Multiple Methods

Víctor Fernández-Carbajales, Miguel Ángel García, José M. Martínez
Grupo de Tratamiento de Imágenes, Escuela Politécnica Superior
Universidad Autónoma de Madrid, E-28049 Madrid, Spain
{ victor.fernandez, miguelangel.garcia, josem.martinez}@uam.es

Abstract

This paper describes and evaluates an algorithm for real-time people detection in video sequences based on the fusion of evidence provided by three simple independent people detectors. Experiments with real video sequences show that the proposed integration-based approach is effective, robust and fast by combining simple algorithms.

1. Introduction

Automatic people detection in video sequences is a complex problem with potential application to a wide variety of tasks mainly related, but not limited, to automated video surveillance and monitoring. The complexity of this problem is mainly due to two reasons. Firstly, there is no easy way to define a model that characterizes how a person will appear in an image, since humans may adopt an enormous variety of poses that can rapidly change, for instance, while walking or even standing. This is worsened by a disguising effect due to clothing, accessories or even hairstyle. Secondly, live video sequences must be analyzed in real time in a majority of applications.

Thus, real time constraints limit the complexity of the detection algorithms that can be applied in practice, while simple and, thereby, fast people detectors cannot independently cope with the intricacy of shapes and appearances that characterize the human body.

This paper proposes a tradeoff solution to that problem by fusing the evidences provided by three relatively simple and fast people detectors. Experiments show that the proposed integrated detector is more effective and robust than the original detectors applied on their own, and does not sacrifice the real-time response required for analysis of live video. These results suggest that such an integration-based approach is likely to be advantageous for combining other people detectors, no matter their complexity, as it allows them to complement each other.

The paper is organized as follows. Section 2 briefly summarizes previous work on people detection, including two of the detectors that have been chosen to be integrated in this work. Section 3 describes how this integration has been performed. Section 4 presents and discusses experimental results with real video sequences, comparing the performance of the proposed integrated detector with the ones of the three original detectors applied on their own. Finally, conclusions are given in section 5.

2. Previous Related Work

People detection algorithms can be classified into two broad families depending on whether they analyze contours (silhouettes) or regions. Most approaches rely on some kind of background segmentation stage that initially extracts relevant objects such that subsequent processing is limited to those parts.

The majority of people detectors currently belong to the contour analysis family. In many cases, they apply a training phase for determining a set of silhouettes representative of typical people poses (e.g., walking, standing, sitting). For instance, [1] creates a table of fixed-length vectors that describe the shape of a set of prototype contours. Those vectors are then compared to the ones extracted from the analyzed images. Alternatively in [2], template silhouettes are compared to the actual contours through the Chamfer's distance. Instead of keeping and comparing representations of real contours, [3] proposes the use of Markov random fields in order to better capture the complexity and variability of silhouettes of people.

A different approach proposed in [6] does not keep full silhouettes but distinctive segments of them referred to as *edgelets*. A training phase searches for edgelets in contours of people, as well as for their co-occurrence probabilities. Those edgelets are then sought in the analyzed images and their co-occurrences evaluated in order to assess the presence or not of people. This method is tolerant to partial occlusions.

The previous techniques rely on the storage and comparison of prototypes of full silhouettes or parts of

them. Alternatively, other schemes are directly based on information extracted from the analyzed silhouettes. For example, the technique in [7] determines the pose and location of body parts from the silhouette of a person. It first applies the Graham's algorithm for computing the convex and concave hulls of the given contour. The vertices of those polygons and the principal directions of the silhouette are used to predict the position of the head and, thereafter, of the hands, feet and torso. Another approach proposed in [8] iteratively computes the largest ellipse contained in every foreground region obtained after background subtraction. The aspect ratio of the ellipses and the iterations required to find them allow the system to determine whether the silhouette corresponds to a person or not. These two techniques have been chosen to be integrated in the algorithm described in this paper for their simplicity and efficiency.

Besides the majority of silhouette-based people detectors, other techniques analyze image regions. For instance, [4] applies color segmentation for extracting uniform regions from the analyzed images. These regions are then matched to the nodes of a tree that represents the constituent parts of a person, as well as their associated topological relationships and probability distributions. Another approach described in [5] considers a model of a person constituted by three rectangles that denote the locations of the torso, head and face. The typical percentages of foreground pixels inside the three rectangles and of skin pixels in the face rectangle are precomputed based on training images. The foreground pixels of the analyzed image are then obtained by background subtraction and a skin color detector is applied. The system then generates multiple hypotheses about the locations of the three rectangles of the model and chooses the ones whose percentages calculated from the analyzed image are most similar to those precomputed for the model.

3. People Detection by Evidence Fusion

The proposed people detector fuses evidences derived from three independent fast people detectors: two of the silhouette-based techniques overviewed in section 2 ([7][8]) and the straightforward aspect ratio. The three detectors are separately applied to every blob detected as a foreground object by a background subtraction scheme proposed in [9]. Occlusions and groups of people are not considered.

A number of people-related features are extracted from each detector after applying it to a specific blob. Each feature is then mapped to a measure of evidence regarding the presence of a person within the blob. The evidences generated in that way from a detector are then adaptively fused in order to generate its overall evidence. Finally, the evidences corresponding to the three detectors with respect to the same blob are adaptively fused into a combined evidence, which is then thresholded in order to make the decision about whether the blob corresponds to a person or not.

In order to generate a measure of evidence from a given people-related feature *x*, the latter is assumed to approximately follow a normal distribution of mean μ and standard deviation σ. Both parameters are experimentally determined for every defined feature by considering a training set with images of people in usual postures. The evidence of the given feature is then defined as a real value between zero and one, the latter when the feature is equal to its associated mean:

$$E_{\mu,\sigma}(x) = e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The first detector being fused is based on a single feature, the aspect ratio of the blob, which is defined as the quotient between its width *w* and height *h*. The first evidence is thus defined as: $E_a = E_{\mu_a,\sigma_a}(w/h)$, where μ_a and σ_a are the mean and standard deviation of the aspect ratios computed from the training set (e.g., $\mu_a = 0.3$, $\sigma_a = 0.2$ in this work).

The second people detector applied to every blob is based on the algorithm proposed in [8], which iteratively computes the largest ellipse contained in the foreground region associated with the blob. All ellipses are sampled into a predefined number of points (e.g., 100). The number of iterations is limited to a maximum (e.g., $i_{MAX} = 100$). The iterative fitting process stops either when the ellipse is fully contained in the blob's foreground region or when the number of iterations reaches the maximum. In the latter case, the ellipse may not be fully contained in the blob's foreground, with some of its points lying outside of it. Three features are evaluated for this detector based on the fitted ellipse:

(*a*) The ratio between the number of iterations and the maximum: i/i_{MAX}. Let μ_{e1} and σ_{e1} be the mean and standard deviation of that ratio as computed from the training set (e.g., $\mu_{e1} = 0.8$, $\sigma_{e1} = 0.15$ in this work). The evidence for this feature is defined as: $E_{e1} = E_{\mu_{e1},\sigma_{e1}}(i/i_{MAX})$.

(*b*) The percentage, *p*, of sampled points that lie outside the blob's foreground, if any. Let μ_{e2} and σ_{e2} respectively be the mean and standard deviation of that percentage as computed from the training set (e.g., $\mu_{e2} = 0.07$, $\sigma_{e2} = 0.12$ in this

work). The evidence corresponding to this feature is defined as: $E_{e2} = E_{\mu_{e2},\sigma_{e2}}(p)$.

(c) The ellipse's aspect ratio, *r*. Analysis of the training set has shown that this feature is distributed as a mixture of two normal distributions of mean and standard deviations (μ_{e3}, σ_{e3}) and (μ_{e4}, σ_{e4}) respectively (e.g., $\mu_{e3} = 0.28$, $\sigma_{e3} = 0.04$, $\mu_{e4} = 0.6$, $\sigma_{e4} = 0.1$ in this work). Hence, the evidence for this feature is defined as: $E_{e3} = E_{\mu_{e3},\sigma_{e3}}(r) + E_{\mu_{e4},\sigma_{e4}}(r)$.

The final evidence, E_e, associated with this second detector is defined as the average of the above three partial evidences: $E_e = (E_{e1} + E_{e2} + E_{e3})/3$.

The third people detector applied to every blob is based on the *Ghost algorithm* proposed in [7], which approximates the contour of the blob's foreground with a closed polygon corresponding to the contour's convex and concave hulls. Three features are also evaluated for this detector based on that polygon:

(*a*) The number of points of the polygon, *n*, which captures the structural complexity of the contour. Let μ_{g1} and σ_{g1} be the mean and standard deviation of that value as computed from the training set ($\mu_{g1} = 27$ and $\sigma_{g1} = 12$ in this work). The evidence for this feature is defined as: $E_{g1} = E_{\mu_{g1},\sigma_{g1}}(n)$.

(*b*) The ratio between the amount of convex and non-convex vertices, *c*, which is usually balanced for a person. Let μ_{g2} and σ_{g2} be the mean and standard deviation of this ratio as computed from the training set (e.g., $\mu_{g2} = 1$, $\sigma_{g2} = 0.5$ in this work). The evidence for this feature is defined as: $E_{g2} = E_{\mu_{g2},\sigma_{g2}}(c)$.

(*c*) The inverse of the number of vertices that are found to belong to the top of the polygon (head) according to the procedure described in [7]: *t*. Let μ_{g3} and σ_{g3} be the mean and standard deviation of this ratio as computed from the training set (e.g., $\mu_{g3} = 0.82$, $\sigma_{g3} = 1.2$ in this work). The evidence for this feature is defined as: $E_{g3} = E_{\mu_{g3},\sigma_{g3}}(t)$.

The final evidence, E_g, associated with the third detector, which is the most complex and robust of the three, is defined as the average of the above three partial evidences, with the provision that the second and third ones are only accounted for if they are above a predefined relevance threshold ρ (e.g., $\rho = 0.6$ in this work):

$$E_g = \frac{E_{g1} + H(E_{g2} - \rho)E_{g2} + H(E_{g3} - \rho)E_{g3}}{1 + H(E_{g2} - \rho) + H(E_{g3} - \rho)}$$

where $H(x)$ is the Heaviside step function.

The final evidence, *E*, about the analyzed blob being a person is obtained by averaging the evidences provided by the three detectors, with the provision that the evidences of the aspect ratio, E_a, and fitted ellipse, E_e, detectors are only accounted for if they are above the predefined relevance threshold ρ:

$$E = \frac{E_g + H(E_a - \rho)E_a + H(E_e - \rho)E_e}{1 + H(E_a - \rho) + H(E_e - \rho)}$$

In the end, the analyzed blob is classified as a person if the combined evidence is above a predefined decision threshold τ (e.g., $\tau = 0.75$ in this work).

4. Experimental Results

The proposed fusion-based people detector has been tested on several public video sequences manually annotated, showing significantly better performance than when the integrated detectors (aspect ratio, fitted ellipses [8] and Ghost [7]) are applied on their own.

For instance, Figure 1 shows the ROC curves (true positive vs. false positive rates) corresponding to the application of the four detectors to the *Hall Monitor* sequence and the *abandoned baggage scenario (stage 1)* from the *i-LIDS* dataset for *AVSS 2007*. Those curves have been obtained by considering three values of the decision threshold τ: 0.6, 0.75 and 0.9.

The proposed scheme is superior to its constituent techniques, yielding the largest rate of true detections with the lowest rate of false detections, especially for the lowest decision threshold (largest true positive rate). In addition, it is far more stable than, for example, the aspect ratio or even the Ghost algorithm, which may both suffer from a significant degradation for low values of τ, even though the evidence of the latter, E_g, is always fused by the proposed scheme.

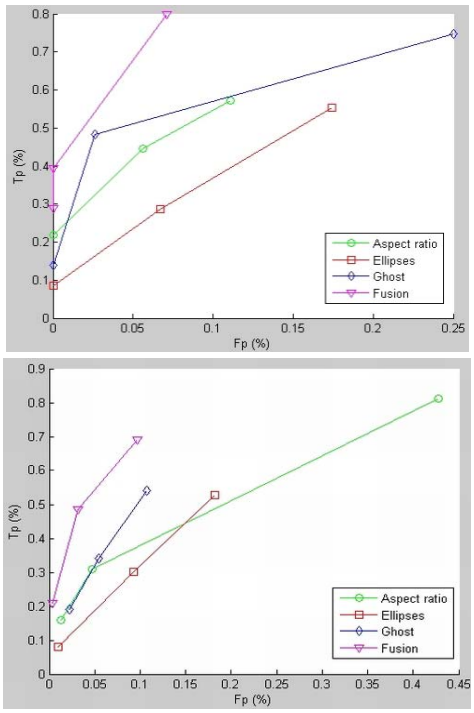

Figure 1. ROC curves for Hall Monitor (top) and AVSS'07 (bottom) video sequences.

Figure 2. (top) Original image and foreground objects. (bottom) Detected people in green.

Figure 2 shows an example of application of the proposed technique to the AVSS sequence. The detection process runs at 3.5 ms in average for every detected blob (Pentium IV @ 3.2 GHz). From this time, 73% corresponds to the Ghost algorithm, 26% to ellipse fitting and less than 1% to the aspect ratio. Further results can be found at "www-gti.ii.uam.es/publications/PeopleDetectionWIAMIS08"

5. Conclusions

A new algorithm for real-time people detection in video sequences has been described and evaluated. The proposed scheme is based on the integration of evidence generated by three independent people detectors. Experimental results show that the proposed scheme is significantly more efficient and stable than when its constituent detectors are applied on their own.

6. Acknowledgements

This work is supported by *Cátedra Infoglobal-UAM para "Nuevas tecnologías de vídeo aplicadas a la seguridad",* the Spanish Government (TEC2007-65400 SemanticVideo) and the *Comunidad de Madrid* (S-050/TIC-0223 - ProMultiDis-CM).

7. References

[1] J.Zhou, and J.Hoang. "Real Time Robust Human Detection and Tracking System", *Proc. of IEEE Computer Vision and Pattern Recognition*, 2005, pp. 149-156.

[2] M.Hessein et al. "Real-Time Human Detection, Tracking, and Verification in Uncontrolled Camera Motion Environments", *Proc. of International Conference on Vision Systems*, 2006, pp. 41-47.

[3] Y.Wu, and T.Yu. "A Field Model for Human Detection and Tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(5):753-765, May 2006.

[4] N.Sprague, and J.Luo. "Clothed People Detection in Still Images", *Proc. of International Conference on Pattern Recognition*, 2002, pp. 585-589.

[5] S. Harasse, and L. Bonnaud. "Human model for people detection in dynamic scenes", *Proc. of IEEE Computer Vision and Pattern Recognition*, 2006, pp. 335-354.

[6] B. Wu, and R. Nevatia. "Detection of Multiple, Partially Occluded Humans in Single Image by Bayesian Combination of Edgeless Part Detector", *Proc. of 10th IEEE International Conference on Computer Vision*, 2005, pp. 90-97.

[7] I. Haritaoglu, D. Harwood, and L.S. Davis. "Ghost: A Human Body Part Labelling System Using Silhouettes", *Proc. of International Conference on Pattern Recognition*, 1998, pp. 77-82.

[8] F.Xu, and D.Fujimura. "Human Detection Using Depth and Gray Images", *Proc. of IEEE Advanced Video and Signal Based Surveillance*, 2003, pp. 115-121.

[9] A.Cavallaro, O.Steiger, and T.Ebrahimi. "Semantic Video Analysis for Adaptive Content Delivery and Automatic Description", *IEEE Transactions of Circuits and Systems for Video Technology*, 15(10):1200-1209, Oct. 2005.