Detecting and Segmenting Humans in Crowded Scenes

Mikel D. Rodriguez
University of Central Florida
4000 Central Florida Blvd
Orlando, Florida, 32816
mikel@cs.ucf.edu

Mubarak Shah
University of Central Florida
4000 Central Florida Blvd
Orlando, Florida, 32816
shah@cs.ucf.edu

ABSTRACT

We describe an approach for detecting and segmenting humans with extensive posture articulations in crowded video sequences. In our method we learn a set of mean posture clusters, and a codebook of local shape distributions for humans in various postures. Detection proceeds in two stages: first instances of the codebook entries cast votes for locations of humans in the video and their respective postures. Subsequently, consistent hypotheses are found as maxima within a voting space. The segmentation of humans in the scene is initialized by the corresponding posture clusters and contours are evolved to obtain precise and consistent segmentations.

Our experimental results indicate that the framework provides a simple yet effective means for aggregating local and global shape-based cues. The proposed method is capable of detecting and segmenting humans in crowded scenes as they perform a diverse set of activities and undergo a wide range of articulations within different contexts.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*; I.4.6 [**Image Processing and Computer Vision**]: Segmentation

General Terms

Algorithms

Keywords

Object Recognition, Human Detection, Segmentation.

1. INTRODUCTION

The ability to accurately detect and segment humans in video sequences represents an essential component in a wide range of application domains such as dynamic scene analysis, human-computer interface design, driver assistance systems, and the development of intelligent environments. Nevertheless, the problem of human detection has numerous challenges associated with it. Effective solutions must be able to account not only for the nearly 250 degrees of freedom of which the human body is capable[14], but also the variability introduced by various factors such as different clothing styles, and the presence of occluding accessories such as backpacks and briefcases. Furthermore, a significant percentage of scenes, such as urban environments, contain substantial amounts of clutter and occlusion.

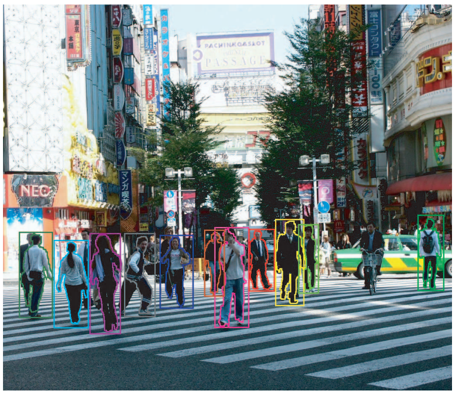

Figure 1: A typical crowded real-world urban scene along with a series of detections from our method.

Despite these challenges, detecting humans within video sequences has constituted an active area of research for a number of years, resulting in the proposal of numerous approaches. Nevertheless, only a small subset of the existing methods has been demonstrated to be effective in the presence of considerable overlaps and partial occlusion in video sequences, such as those seen in Figure 1.

The shape of the human silhouette is often very different from the shape of other objects in a scene. Therefore, shape-based detection of humans represents a powerful cue which can be integrated into existing lines of research. As opposed to the appearance-based models, human shapes tend to be somewhat isotropic. Hence, shape-based methods coupled with other cues, such as motion, can provide a discriminating factor for recognition.

In this paper we address the challenge of detecting and segmenting humans in video sequences containing crowded real-world scenes. Due to the difficulty of this problem, reliance on any single global model or feature alone would be

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009 ...$5.00.

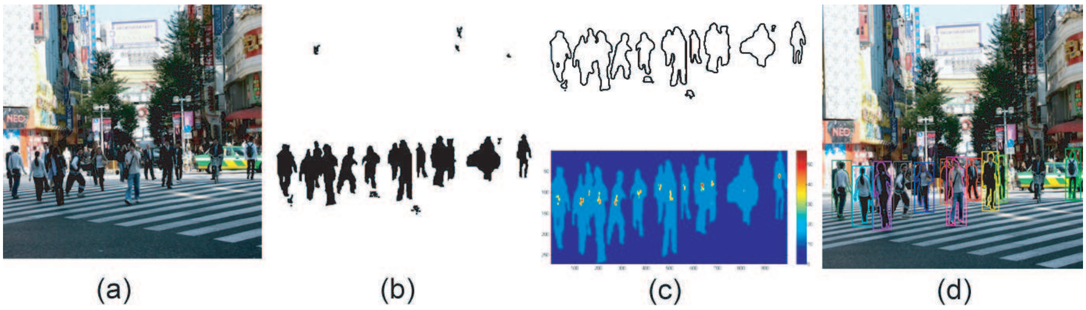

Figure 2: The main steps in human detection. (a) Input frame, (b) Foreground blobs, (c) Contours extracted from the foreground blobs are depicted at the top. The bottom row depicts the votes cast by codebook instances, representing hypotheses of human centroids in the scene, (d) Final detection and segmentation.

ineffective. Therefore, successful approaches must be capable of integrating both global and local shape cues.

2. RELATED WORK

Currently, the most prevalent class of methods present in the literature is the detector-style method, in which detectors are trained to search for humans within a video sequence over a range of scales. A number of these methods use *global* features such as edge templates [3], while others build classifiers based on *local* features such as SIFT-like descriptors [6], Haar wavelets [11] and the histogram of oriented gradient descriptor [2].

Another family of approaches models humans as a collection of parts [5], [7], [8] and [9]. Typically this class of approaches relies on a set of low-level features which produce a series of part location hypotheses. Subsequently, inferences are made with respect to the best assembly of existing part hypotheses. Approaches such as AdaBoost have been used with some degree of success to learn body part detectors such as the face [10], hands, arms, legs, and torso [5] [8].While this class of approaches is attractive, detection of parts is itself a challenging task. This is particularly difficult in the class of scenes in which we are interested, which consist of crowded scenes containing significant occlusion amongst many parts.

A considerable amount of work has also focused on shape-based detection. Zhao et al [16] use a neural network that is trained on human silhouettes to verify whether the extracted silhouettes correspond to a human subject. However, a potential disadvantage of the approach resides in the fact that they rely on depth data to extract the silhouettes. Our work is similar in principle to the framework presented in [4], in which a patch-based approach is used to learn an implicit shape model for walking humans. Others, such as Davis et al [15] have also attempted to make use of shape-based cues by comparing edges to a series of learned models. Wu et al [12] have proposed learning human shape models and representing them via a Boltzmann distribution in a Markov Field. Although a number of these methods have proved to be successful in detecting humans in still images, most of them assume isolated human subjects with a minimal presence of clutter and occlusion. We are interested in methods that enable effective use of shape-based cues in the presence of heavy occlusion of the kind present in most urban environments.

3. APPROACH

In this section we describe our approach for detecting and segmenting humans in cluttered video sequences. The method is based on a formulation that integrates local shape distributions and mean posture contours.

The method begins by learning a set of global posture clusters which are used to initialize segmentation. Additionally, we learn a codebook of local shape distributions based on humans in the training set. When the system is presented with a new frame from the testing video sequence, it extracts contours from the foreground blobs in each frame, samples them using shape context, finds instances of the learned local shape codebook, and casts votes for human locations and their respective postures in the frame. Subsequently, the system searches for consistent hypotheses by finding maxima within the voting space. Given the locations and postures of humans in the scene, the method proceeds to segment each subject. This is achieved by projecting the mean posture shape corresponding to the posture cluster of every consistent hypothesis around the centroid vote.

3.1 Learning

Given a a set of training videos, we perform background subtraction and extract contours from each frame by performing edge detection on foreground blobs corresponding to humans in the scene. Each contour is sampled at d points (in our experiments d ranges from 1000-3000). For a point p_i on the contour, we identify a distribution over relative positions by computing a coarse histogram h_i of the relative coordinates of the $d-1$ points that remain. These histograms are concatenated into a single shape vector for every silhouette. Shape vectors are then clustered into n clusters using K-means. In our experiments n was set to 150, these cluster centers represent a set of typical global human shapes in the training set (Figure 3).

Once the posture clusters are created the second phase of the process consists of learning a codebook of local shapes and their spatial distribution for different posture clusters. The intuition behind this lies in the fact that in crowded scenes reliance on global shape models alone would likely be ineffective. Therefore, we create a codebook of local shapes by sampling points on silhouettes using shape context descriptors [1]. Subsequently, all of the shape context descriptors are clustered into M clusters using K-means. The similarity of the shape descriptors is given by the χ^2 distance

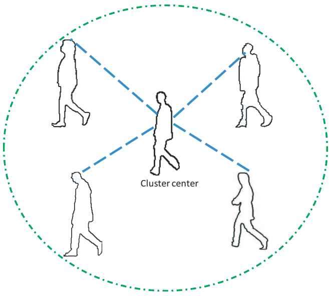

Figure 3: Each posture cluster is represented by a mean contour (depicted here in the center). The collection of all posture cluster centers represent the set of typical global human shapes in the training data.

between the two K-bin histograms $g(k)$ and $h(k)$. χ^2 is given by:

$$\chi^2 = \frac{1}{2}\sum_{k=1}^{K}\frac{[g(k)-h(k)]^2}{g(k)+h(k)}. \quad (1)$$

For each of the M clusters we store the cluster center as a codebook entry (in our experiments the size of the codebook was set to 400). We then learn the spatial distribution of the codebook entries for different postures. This is done by iterating through all of the foreground blobs from the training set, sampling each silhouette via shape context descriptors, and matching these against codebook entries. For each instance of a codebook entry we record two pieces of information: The position with respect to the centroid of the human silhouette on which it occurred, and the closest posture cluster to which the silhouette belongs.

3.2 Detection and segmentation

Given a testing video sequence, at each frame we extract contours from the foreground blobs produced by background subtraction. These contours are then sampled using shape context, producing a series of shape context descriptors for each contour. Each descriptor is then compared to the learned codebook of local shapes. If a match is found, the corresponding codebook entry will cast votes for the possible centroid of a human in the scene (Figure 4) and a posture cluster to which it belongs. Votes are aggregated in a voting space and Mean-Shift is used to find maximums. The various steps associated with detection are depicted in Figure 2.

Let $\mathbf{s}$ be our local shape observed at location l. If $\mathbf{s}$ matches to a set of codebook entries C_i, each activation C_i will be weighted by the quality of the match given by the χ^2 distance between $\mathbf{s}$ and C_i. Each codebook activation will cast a set of votes for the possible centroid x and posture p_n of humans within the scene.

Given a set of hypotheses H_i corresponding to (p_n, x), we search for consistent votes by integrating hypotheses within a search window using Mean-Shift.

After selecting the strongest set of hypotheses for human locations and their postures contours are initialized by projecting the silhouettes associated with the posture cluster p_n at location x in the video frame. Subsequently, contours are evolved using a level-set based segmentation approach proposed in [13]. Both texture and color features are used in order to guide the segmentation of humans present in the scene (Figure 5).

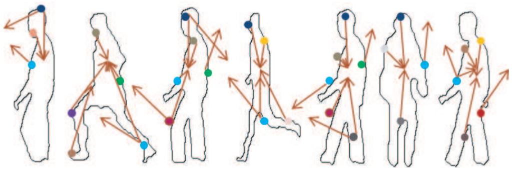

Figure 4: Local shape codebook instances (which are depicted as small colored dots on the contour) vote for the location of humans within the scene.

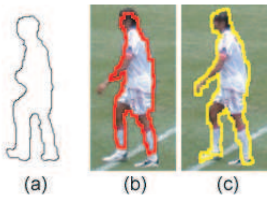

Figure 5: Given a hypothesis for x and h_n, segmentation is initialized with the stores posture cluster silhouette. (a) The posture cluster silhouette to which the subject on the right belongs. (b) Segmentation based only on foreground pixels results in a inaccurate segmentation. (c) When the segmentation is initialized by the posture cluster silhouette and contours are evolved we obtain improved precision and consistency.

4. EXPERIMENTS AND RESULTS

We evaluated the performance our system for detecting and segmenting humans on a set of challenging video sequences containing significant amounts of partial occlusion. Furthermore, the videos included in the testing procedure featured humans performing a diverse set of activities within different contexts, such as walking on a busy city street, running a marathon, playing soccer, and participating in a crowded festival.

Our training set ranged from 700 to 1,100 frames from the various video sequences containing human samples. Metadata in the training set included the centroid of each silhouette, along with the posture cluster to which it belonged. Our testing database consisted of a wide range of scenes, totalling 34,100 frames in size and contained a total of 312 humans for which the torso is visible. The size of the humans across the video sequences averaged 22x52 pixels. Figure 6 shows some examples from the data set.

The quantitative analysis of the method centered around measuring correct detection as well as the reported locations of humans within the scene. We employed the evaluation framework proposed in [4], which consist of three criteria: The first criteria, *relative distance*, measures the distance between bounding box centers with respect to the size of the ground truth box. The second and third criteria (*cover* and *overlap*) measure the common area between the detected

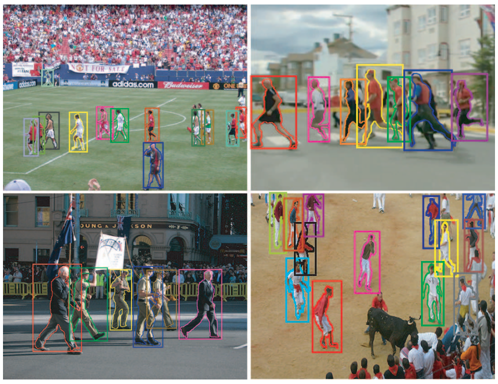

Figure 6: Example detections from the testing set are depicted by a colored bounding box and posture-based segmentations.

bounding box and the ground truth. In our experiments a detection is classified as being correct if the deviation in relative distance is less than 25% and less then 50% for cover an overlap.

In addition to evaluating the detection results of the method we also assessed the effect of the training set size. This was achieved by varying the number of frames used to learn the spatial distribution of local shapes. Figure 7 depicts the receiver operating characteristic (ROC) curve for the different modes.

Our detection results show that the method achieves high recognition rates and performs reliable segmentations, despite the presence of significant occlusions in the scene. On the most challenging video sequence the system achieved on average a 75.3% recognition rate. Whereas on the best video we achieved an recognition rate of 94%. The false positive rate was low, most of the erroneous detections were caused by moving vertical structures that resembled the human shape. Given the difficulty of the data set, these results are encouraging.

As demonstrated by Figure 7, the effect of the size of the training set on the performance of the method was minimal. Although an overall increase in performance is observed with a larger training set, a reduction of the training set size does not lead to a drastic decrease in the performance of the method.

5. CONCLUSIONS

We have presented a framework for detecting and segmenting humans in real-world crowded scenes which integrates both local and global shape cues. Cluttered scenes containing many occlusions render the lone use of global shape representations as ineffective. Instead we aggregate local shape evidence via a codebook of local shape distributions for humans in various postures. Additionally, we found that a set of learned global posture clusters aids the segmentation process.Our experiments indicate that local shape distribution represents a powerful cue which can be integrated into existing lines of research.

6. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[2] N. Dalai, B. Triggs, I. Rhone-Alps, and F. Montbonnot. Histograms of oriented gradients for human detection. *CVPR*, 1, 2005.

[3] D. Gavrila. Pedestrian detection from a moving vehicle. *ECCV*, 2:37–49, 2000.

[4] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR*, 1, 2005.

[5] A. Micilotta, E. Ong, and R. Bowden. Detection and Tracking of Humans by Probabilistic Body Part Assembly. *BMVC*, 2005.

[6] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, 1:69–81, 2004.

[7] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: tracking people by finding stylized poses. *CVPR*, 1, 2005.

[8] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. *ECCV*, 4:291–303, 2004.

[9] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. *ECCV*, 4:700–714, 2002.

[10] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 1:511–518, 2001.

[11] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. *IJCV*, 63(2):153–161, 2005.

[12] Y. Wu and T. Yu. A Field Model for Human Detection and Tracking. *CVPR*, 28, 2006.

[13] X. Yilmaz, A. Li and Shah. Contour-Based Object Tracking with Occlusion Handling in Video Acquired Using Mobile Cameras. *PAMI*, 26(11):1531–1536, 2004.

[14] V. Zatsiorsky. *Kinetics of Human Motion*. Human Kinetics, 2002.

[15] L. Zhao and L. Davis. Closely Coupled Object Detection and Segmentation. *ICCV*, 1, 2005.

[16] L. Zhao and C. Thorpe. Stereo-and neural network-based pedestrian detection. *ITS, IEEE Transactions on*, 1(3):148–154, 2000.

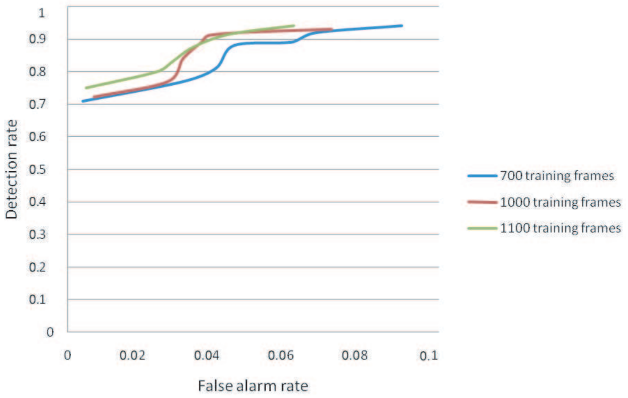

Figure 7: Recognition performance based on a range of training set sizes.