COARSE-TO-FINE PEDESTRIAN LOCALIZATION AND SILHOUETTE EXTRACTION FOR THE GAIT CHALLENGE DATA SETS

Haiping Lu, K.N. Plataniotis and A.N. Venetsanopoulos

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto, M5S 3G4, Canada
{haiping, kostas, anv}@dsp.toronto.edu

ABSTRACT

This paper presents a localized coarse-to-fine algorithm for efficient and accurate pedestrian localization and silhouette extraction for the Gait Challenge data sets. The coarse detection phase is simple and fast. It locates the target quickly based on temporal differences and some knowledge on the human target. Based on this coarse detection, the fine detection phase applies a robust background subtraction algorithm to the coarse target regions and the detection obtained is further processed to produce the final results. This algorithm has been tested on 285 outdoor sequences from the Gait Challenge data sets, with wide variety of capture conditions. The pedestrian targets are localized very well and silhouettes extracted resemble the manually labeled silhouettes closely.

1. INTRODUCTION

Gait recognition [2], the identification of individuals in video sequences by the way they walk, has recently gained significant attention. This interest is strongly motivated by the need for automated person identification system at a distance in visual surveillance and monitoring applications in security-sensitive environments such as banks and airports, where other biometrics such as fingerprint, face or iris information are not available at high enough resolution for recognition [3]. In [4], Sakar *et al.* introduced the HumanID Gait Challenge problem, providing a set of twelve experiments of increasing difficulty, which examine the impact of five covariates on performance. The challenge provided the means to measure progress in the area and various researchers have reported results on these data sets [4, 5, 6]. However, most of them are using silhouettes obtained semi-automatically with manual outlining of bounding boxes [4]. In [5], the silhouettes are extracted automatically but under the assumption that the paths of the silhouette centroid must be smooth to a second degree polynomial. Moreover, the Gait Challenge data sets include difficult sequences with noise resulting from heavy shadows, camouflaging effects, other subjects in the scene, etc. Automatic handling of these difficulties is important to the advancement of the gait recognition research.

There are a number of background subtraction algorithms available but most of them are pixel-wise processing [7]. In [1], a novel background subtraction algorithm using Markov thresholds is proposed. This method extracts the silhouettes of moving objects from a stationary background using Markov random fields (MRF) of binary segmentation variates so that the spatial and temporal dependencies imposed by moving objects on their images are exploited. It is shown that their method produces more accurate and visually appealing silhouettes that are less prone to noise and background camouflaging effects than traditional per-pixel based methods. Three MRFs were proposed with increasing complexity. However, since an annealing procedure is needed, it is a costly and slow algorithm, especially when applied to high-resolution full size color sequences, such as those in the Gait Challenge data sets.

In this paper, a coarse-to-fine approach is proposed for automatic pedestrian localization and silhouette extraction for the Gait Challenge data sets, where only one pedestrian in the view field is of interest. The coarse detection phase is simple and fast to localize the subject roughly and the fine detection phase applies the background subtraction using Markov thresholds (BSMT) algorithm [1] to get an accurate estimation of the silhouettes. Domain knowledge (e.g., on the shape and motion of the pedestrian) is incorporated to produce robust detection results. Experiments show robust pedestrian localization results and improved silhouette extraction by evaluation against manually labeled silhouettes (as the ground truth), compared with the methods in [4] and [5].

2. THE PROPOSED ALGORITHM

The proposed pedestrian localization and silhouette extraction algorithm consists of two phases, as shown in Fig. 1.

The authors would like to thank J. Migdal from the MIT for providing the source codes of their algorithm in [1]. The authors would also like to thank Prof. S. Sarkar from the USF for providing the manual silhouettes and the Gait Challenge data sets. Support provided by the Communications and Information Technology Ontario Partnership Program and the Bell University Labs - at the University of Toronto is also acknowledged.

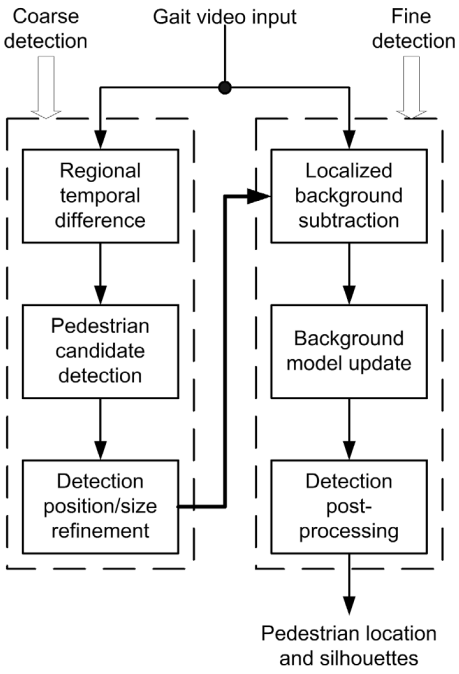
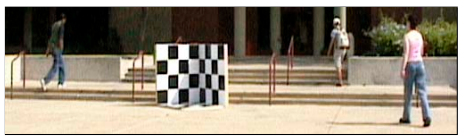


Fig. 1. The coarse-to-fine pedestrian localization and silhouette extraction algorithm.

2.1. Silhouette extraction difficulties

The Gait Challenge data sets are captured under various outdoor conditions. Since they are outdoor data sets, the existence of other pedestrians, slow-motion of pedestrians, heavy shadows and flattering construction strips impose significant difficulties in successful extraction of the pedestrian subject. Fig. 2 shows some examples of the difficulties.

(a) Heavy shadow.



(b) Other subjects in the scene.

(c) Slow motion (eight successive frames shown).

Fig. 2. Examples of difficulties in pedestrian silhouette extraction for the Gait Challenge data sets.

Background subtraction algorithms [7] are commonly used to generate silhouettes. In [4], bounding boxes around the moving person are defined semiautomatically in each frame of a sequence, and silhouettes are then extracted by adaptively deciding on the foreground and background labels for each frame by estimating the foreground and background likelihood distributions using the iterative expectation maximization (EM) procedure, with Gaussian Mixture Model (GMM) for the observations. In [5], the walking path is assumed to be smooth to a second degree polynomial, and bounding boxes are obtained through repeated robust estimation, with Gaussian model for background modeling. The algorithm proposed in this paper is fully automatic and there is no specific assumption made regarding the walking path of the pedestrians. Furthermore, successful silhouette extraction by background subtraction is not always possible, especially for the first a number of frames (since it takes time to learn the background model) and in the case of slow motion of the subject. Thus, pedestrian localization is another objective besides silhouette extraction so that other algorithms such as model-based solutions [8] can be applied in the case of silhouette extraction failure.

2.2. Coarse detection: simple and fast

When a new frame arrives, the proposed algorithm first detects its foreground pixels in a coarse region $\mathcal{R}_c$ centered at the previous fine detection box $\mathcal{B}_f$ with an offset of α pixels at each side.

A gray map $\mathcal{M}_1$ is used to record the maximum pixel differences (across the color channels: red, green and blue). Foreground pixels $\mathcal{F}_1$ are detected by thresholding the simple frame differences with threshold $\mathcal{T}_d$.

If the number of foreground pixels detected exceeds a threshold $\mathcal{T}_{n1}$, the spatial distribution of these pixels is examined. Pedestrian silhouettes to be extracted are large objects with a roughly rectangular shape. Therefore, the human subject is localized by searching for a rectangular box enclosing sufficiently large number ($\mathcal{T}_{n2}$) of foreground pixels $\mathcal{F}_1$, with the left-top corner of the rectangular region being a foreground pixel (in $\mathcal{F}_1$) with at least two connected pixels in $\mathcal{F}_1$. All pixels in the rectangular box found are labeled as foreground pixels in $\mathcal{F}_2$, resulting in a binary map $\mathcal{M}_2$.

Next, a bounding box is obtained for the target pedestrian. For foreground pixels ($\mathcal{F}_2$) in $\mathcal{M}_2$, the box width and height are obtained from the maximum and minimum row and column indexes and the box center is determined by the centroid (x_c, y_c) of the foreground pixels in $\mathcal{M}_2$, with pixel frame differences in $\mathcal{M}_1$ as the weight:

$$x_c = \frac{\sum_{(x,y)\in\mathcal{F}_2} x \cdot \mathcal{M}_1(x,y)}{\sum_{(x,y)\in\mathcal{F}_2} \mathcal{M}_1(x,y)}, \quad (1)$$

$$y_c = \frac{\sum_{(x,y)\in\mathcal{F}_2} y \cdot \mathcal{M}_1(x,y)}{\sum_{(x,y)\in\mathcal{F}_2} \mathcal{M}_1(x,y)}. \quad (2)$$

Through the weighting, the bounding box center is biased towards locations with larger pixel frame differences that are more confident to be true foreground pixels. A pedestrian typically results in a box with the height greater than the width,

thus, only those boxes satisfying this constraint are considered as valid detection. When both the current detection and previous detection are valid, the variation in box width and height, and the changes of the four box side positions are limited to a small number, since silhouette sizes and positions are expected to vary gradually in pedestrian walking. The output of this coarse detection process is a coarse box $\mathcal{B}_c$.

If the current detection is invalid (detection failure), the current coarse detection box $\mathcal{B}_c^t$ is set to $\mathcal{B}_c^{t-1}$, where the superscript is the time index. This is especially useful when the cause of detection failure is slow motion of the subject.

2.3. Fine detection: robust and accurate

The outputs of this procedure are the centered fine detection region $\mathcal{R}_f^c$, the fine detection box $\mathcal{B}_f$, and the silhouette $\mathcal{S}_f$.

The BSMT algorithm is applied to $\mathcal{R}_f$, which is centered at the detected coarse bounding box $\mathcal{B}_c$, with offset of β pixels ($\beta < \alpha$), to get a raw silhouette $\mathcal{S}_r$. The background model is updated using the Gaussian mixture model [9], with the pixels outside $\mathcal{R}_f$ considered as all background pixels and the pixels inside $\mathcal{R}_f$ according to the BSMT results.

The resulted silhouette $\mathcal{S}_r$ is further processed to obtain $\mathcal{B}_f$ and $\mathcal{S}_f$. The vertical and horizontal projections of $\mathcal{S}_r$ are obtained. From the horizontal projection, the top (minimum row) and bottom (maximum row) of $\mathcal{B}_f$ are determined. Next, based on connected region analysis of the vertical projections, the silhouette $\mathcal{S}_r$ is separated into clusters, and the fine detection bounding box $\mathcal{B}_f$ of the pedestrian corresponds to the cluster with the maximum vertical projection. The region $\mathcal{R}_f$ is then horizontally re-centered at the foreground horizontal centroid of the silhouette $\mathcal{S}_r$ to get $\mathcal{R}_f^c$. The final silhouette extracted $\mathcal{S}_f$ is the portion of $\mathcal{S}_r$ that is within $\mathcal{R}_f^c$.

2.4. Initial detection

At the beginning of one gait sequence, no knowledge is available about the subject's whereabouts. Therefore, for the first a few frames, the coarse detection procedure is skipped and $\mathcal{R}_f$ is set to be the whole frame in the fine detection procedure until the number of foreground pixels in the fine detection box $\mathcal{B}_f$ exceeds some threshold (e.g., 50) so that it is confident that a pedestrian is localized well in the frame. Thus, the proposed algorithm has a "slow-start" feature.

3. EXPERIMENTAL RESULTS

The proposed algorithm is tested on 285 sequences from the five Gait Challenge data sets (the gallery and probes B, D, H and K), with an average of 630 frames in each sequence. Each frame is a color (RGB) image of size 480×720 and all the following parameters are determined experimentally, with reference to the recommendations in [1] for background subtraction. The rectangular box searched in step 2 of the coarse detection is of size 100×50, and $\alpha = 100$, $\beta = 25$, $\mathcal{T}_d = 15$, $T_{n1} = 100$ and $T_{n2} = 500$. The sizes of the bounding boxes ($\mathcal{B}_c$, $\mathcal{B}_f$) are bounded by a maximum of 250×150 and minimum of 100×50. The background is modeled by 3 Gaussian mixtures [1, 9], with the weight learning rate set to 0.005 and the Gaussian learning rate set to 0.05. In the BSMT algorithm, the M_1 structure is used, with 30 iterations in the annealing procedure.

3.1. Pedestrian localization results

Since it takes time to learn the background model, the target subject is not expected to be localized well at the beginning of a sequence. Thus, the localization performance is evaluated on frames after it is confident that the subject is located well, determined by the number of foreground pixels in the fine detection bounding box $\mathcal{B}_f$.

On average, approximately 50 frames are needed to localize the pedestrian well (i.e., to gain confidence on the subject's whereabouts). Denote the number of foreground pixels in $\mathcal{S}_r$ as F_s and define the dislocation D as

$$D = \frac{|X_{cs}^{R_f} - W_{R_f}/2|}{W_{R_f}/2}, \tag{3}$$

where $X_{cs}^{R_f}$ is the foreground horizontal centroid of $\mathcal{S}_r$ (with the origin at the left top corner of $\mathcal{R}_f$) and W_{R_f} is the width of $\mathcal{R}_f$ (so that $W_{R_f}/2$ is the horizontal center of $\mathcal{R}_f$). An error is logged if $F_s < 50$ or $D > 0.25$ (these threshold values are determined through visual examination of the results). Experiments show that only 117 ($\approx 0.07\%$) out of $165,749$ frames are in error. Furthermore, visual examination of the results (especially frames in error) shows that the pedestrian subjects are all located well except that in a few frames, some (mostly lower) portions of the pedestrian bodies are cut and missing, and sometimes only small portions of the complete silhouettes can be obtained through background subtraction.

As mentioned before, pedestrian localization in case of silhouette extraction failure is useful to further processing for better silhouette extraction, e.g. by employing human body model [8] and appearance modeling [10, 11, 12].

3.2. Silhouette extraction results

The silhouette extraction performance is evaluated by measuring the resemblance between the extracted silhouettes and the corresponding manually labeled silhouettes in [13], with a total number of 10005 frames available. The metric used measures the ratio of the intersection of the silhouettes to the union of the silhouettes:

$$R(A, B) = \frac{A \bigcap B}{A \bigcup B}, \tag{4}$$

where A and B are binary segmentations (silhouettes). This metric is also called the Tanimoto similarity measure and used in [4]. The performance comparison is shown in Fig. 3.

In Fig. 3, number 1 on the horizontal axis represents the Gallery set and numbers 2 to 5 represent probes B, D, H and K, respectively. The last number 6 represents the average of these five sets. The vertical axis represents the average of the metric (4), measured against the manually labeled silhouettes in [13]. The figure shows the results obtained by the semi-automatic methods of USF (USF Semi-Auto), the proposed coarse-to-fine algorithm (UT Coarse-to-Fine) and the silhouette refinement algorithm of MIT (MIT Sil-Refine). The silhouettes for probes H and K are not available for the MIT Sil-Refine algorithm. The proposed algorithm (UT Coarse-to-Fine) is observed to have consistently better performance than the other two methods.

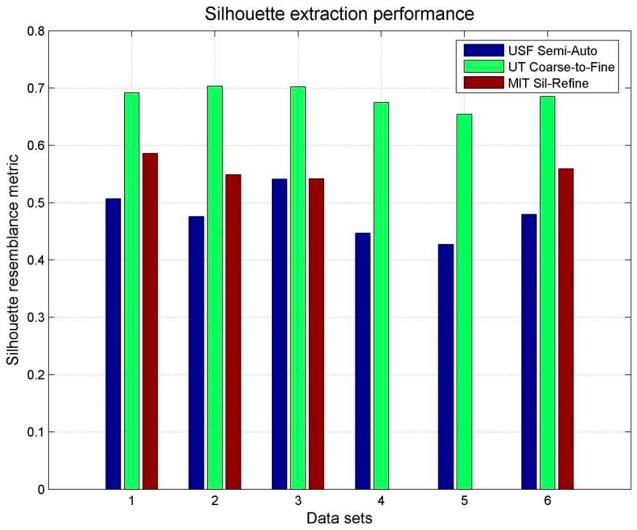


Fig. 3. Performance comparison through resemblance with the manually labeled silhouettes.

4. CONCLUSIONS AND FUTURE WORK

Recently, gait recognition has attracted much attention for its potential to surveillance and security applications. The release of the Gait Challenge data sets provides a common database for testing and evaluation of gait recognition algorithms. The difficulties in the data sets include noise resulting from slow motion of subjects, heavy shadow, other moving subjects and objects in the scene.

This paper proposes a coarse-to-fine automatic pedestrian localization and silhouette extraction algorithm. The coarse detection phase quickly locates the subject through frame differencing and thresholding. The fine detection phase applies a robust background subtraction (using Markov threshold) algorithm [1] to get a more accurate detection, with further post-processing to refine the results. Experiments show that with localized coarse-to-fine processing, the proposed algorithm achieves robust localization results and the silhouettes extracted resemble better with the manually labeled silhouettes (ground truth), compared with two algorithms in the literature.

Future work includes applying human body model such as the layered deformable model developed recently [8] and appearance modeling [10, 11, 12] to help silhouette extraction, especially in detection failure handling.

5. REFERENCES

[1] Joshua Migdal and W. Eric L. Grimson, "Background subtraction using markov thresholds," in *IEEE Workshop on Motion and Video Computing*, January 2005.

[2] M. S. Nixon and J. N. Carter, "Advances in automatic gait recognition," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, May 2004, pp. 139–144.

[3] A. Kale, A. N. Rajagopalan, A. Sunderesan, N. Cuntoor, A. Roy-Chowdhury, V. Krueger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. Image Processing*, vol. 13, no. 9, pp. 1163–1173, Sept. 2004.

[4] S. Sarkar, P. J. Phillips, Z. Liu, I. Robledo, P. Grother, and K. W. Bowyer, "The human ID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[5] L. Lee, G. Dalley, and K. Tieu, "Learning pedestrian models for silhouette refinement," in *Proc. IEEE Conf. on Computer Vision*, Oct. 2003, pp. 663–670.

[6] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: a challenging signal processing technology for biometrics," *IEEE Signal Processing Mag.*, vol. 22, no. 6, Nov. 2005.

[7] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, Oct. 2004, vol. 4, pp. 3099–3104.

[8] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A layered deformable model for gait analysis," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Apr. 2006, pp. 249–254.

[9] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2003.

[10] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 10, pp. 1296–1311, Oct. 2003.

[11] B. J. Frey, N. Jojic, and A. Kannan, "Learning appearance and transparency manifolds of occluded objects in layers," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2003, vol. 1, pp. 45–52.

[12] J. Lim and K. Kriegman, "Tracking humans using prior and learned representations of shape and appearance," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, May 2004, pp. 869–874.

[13] Z. Liu, L. Malave, and S. Sarkar, "Studies on silhouette quality and gait recognition," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 27 June-2 July 2004 2004, vol. 2, pp. 704–711.