Comparative evaluation of two multisensory video surveillance techniques for pedestrian tracking

Girija Chetty
Faculty of Information Sciences and Engineering,
University of Canberra, Australia
girija.chetty@canberra.edu.au

Abstract

In this paper we examine two different automated video surveillance techniques for detection and tracking of pedestrians based on fusion of colour and thermal images. The first approach is a novel particle filtering based on Bayesian framework, and the seconf one is an approach based on fusion of shape and appearance cues. The shape and appearance based technique involves a layered two pass scheme, where in the first pass-an expectation-maximization (EM) algorithm is used to separate infrared images into still background and moving foreground layers. In the second pass: shape cues from the first pass is used to eliminate non-pedestrian moving objects and then appearance cue is used to locate the exact position of pedestrians. Then pedestrians are detected by sequential application of templates at multiple scales. For tracking the pedestrian a graph matching-based algorithm which fuses the shape and appearance information was used. The particle filtering based algorithm on other hand is based on building a scene background model with each pixel represented as a multimodal distribution of colour and thermal images. Then this background model is used to build a particle filter for tracking the pedestrian. The particle filter uses a novel formulation of observation likelihoods The evaluation of the two detection and tracking approaches was done by performing experiments on the thermal and colour dataset from OTCBVS database [1, 2].

1. Introduction

Automated video surveillance involves real-time observation of people and vehicles within a busy environment, leading to a description of their actions and interactions. The technical challenges that characterize this domain include moving object detection and tracking such as pedestrians, object classification, human motion analysis, activity understanding, etc. The problem of pedestrian detection and tracking is a classical computer vision problem with various algorithms proposed for addressing the challenges such as large variability of body pose, clothing and operating environments [3, 4,5,7,8,9.10,11]. The wavelet based appearance representations with support vector machine (SVM) classifier were proposed in [3,4]. In [5-7], silhouette- and shape-based detection techniques was adopted. In [8–11], a joint modeling of human body, pose and motion was done for detecting pedestrians. In [12], periodicity and self-similarity of human motion analysis to detect pedestrians is proposed. An AdaBoost classifier for pedestrian recognition involving appearance and motion based feature vectors is used in [2].An approach based on joint use of PCA vectors and time-delay neural networks for object recognition and tracking is proposed n [13]. Finally, a a stereo-based disparity segmentation and neural network-based pedestrian recognition algorithm was used in [14].

However, most of these approaches used are based on detection and tracking of pedestrians in visible spectrum. Under difficult operating environments (e.g., in nights or bad weathers), sensing in visible spectrum becomes infeasible or severely impaired, which calls for the imaging modalities beyond visible spectrum. With dramatic reduction in cost of thermal sensors in the past few years, it has become possible to deploy infrared (IR) sensors with high dynamic range and sensitivity in applications such as night-vision and all weather surveillance. Several approaches using infra-red imagery has been proposed recently In [15], probabilistic templates were used to capture the variations in human shape for pedestrian detection. A support vector machine with Kalman filter based

978-1-4244-4242-3/08/$25.00 ©2008 IEEE

pedestrian detection and tracking is used in based In [16]. In [17], the P-tile method was proposed to detect human head first, and then human torso and legs are included by local search. In [18], human detection in IR imagery using a particle swarm optimization based algorithm was proposed. In [19], a two-stage template-based method with an Adaboost classifier was presented for pedestrian detection

Few approaches consisting of several cameras were proposed for achieving 24-7 continuous monitoring. Also, some IR surveillance systems based on non-imaging measurements were also proposed [7]. A project named "Smart Floor" [7] aims to identify and track a user around the space with force measuring load cells installed under the floor.

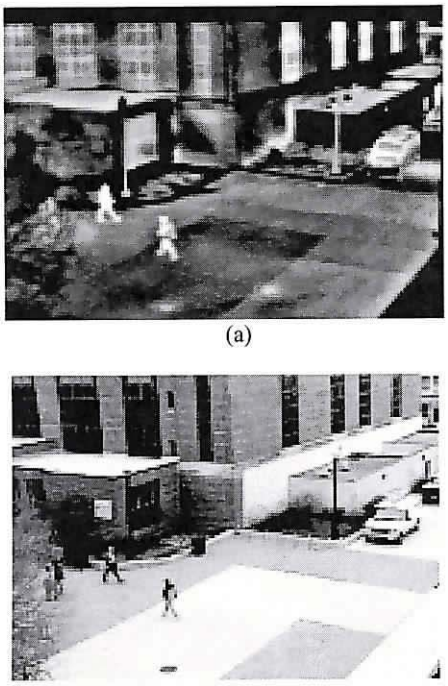

Fig. 1.(a): OTSBVS dataset sample: Thermal image of the scene from the OTS; *(b):* Colour image of the same scene. [3]

For the scenarios cover overlapping areas, a fusion of multiple camera sensors can be used to obtain a more accurate and more efficient solution. Sensor fusion has become an increasingly important direction in computer vision and in particular human tracking systems in recent years. One of the most significant work in this area is a Bayesian approach based on particle filters proposed in [19], where a fusion of the colour camera input with a thermal camera input is used , producing for each pixel a gray scale mapping of the temperature at the corresponding location. Another significant work based on the recognition paradigm is proposed in [20], which processes the fusion of colour and thermal camera inputs using a joint shape and appearance based algorithm.

In this paper we compare the two approaches [19,20] based on a common operational framework using colour and thermal dataset of the open-source OTCBVS [1,2] benchmark pedestrian dataset collection. Figure 1 show a sample data from the database. The paper is organized as follows. Next section describes the layered shape and appearance based approach [20]. The particle filter based approach described in [19] is outlined in Section 3. The experiments carried out on the Dataset 03 for the OTCBVS database for two approaches is described in Section 4, and the paper concludes with some conclusions and plan for further work in Section 5.

2. Layered Shape and Appearance Based Approach

This approach consists of three main stages:

- *Representation of layers*: For representing layers the IR image is decomposed into two layers: background (still objects) and foreground (moving objects). Then a global motion model is used for registering the background images.
- *Joint shape-and-appearance detection:* In this stage the pedestrians are detected from the foreground layer by jointly exploiting the shape and appearance information in two steps: (1) Shape-based classification, (2) Appearance-based localization. In the classification step, an SVM is trained to extract the compactness and leanness of the pedestrian object Then a multistage PCA is used to localize the pedestrians with various sizes in the foreground layer. This joint shape and appearance based formulations allows better performance for the crowded pedestrian situations for low-SNR IR imagery.
- *Shot segmentation and tracking:* In this stage the pedestrian tracking for long sequences is done by addressing the shot segmentation problem. The sequence is first segmented into shots (temporally correlated frames) based on Hausdorff distance; then within each shot, pedestrian tracking is formulated as a matching

problem on weighted bipartite graphs. Each pedestrian corresponds to a node and every potential matching between two nodes in adjacent frames to a weighted edge whose weight reflects the tradeoff between shape-appearance similarity and geometric proximity. Further details of this technique is described in [20].

3. Particle Filtering Based Approach

Particle filters use a novel Bayesian formulation and do not make Gaussianity assumptions [15, 16]. They have potential to perform better in resolving ambiguities while dealing with crowded environments [13. 16]. In this approach, first a dynamically adapting background model is used to segment the foreground regions Then, a head-candidate selection algorithm is used to hypothesize the number of human bodies in the foreground region. As the next step, a Bayesian inference model is constructed by using the priori knowledge of the human parameters, the scene layout and geometry. Further, observations of the body appearances at each frame are used in the probabilistic scheme.

For background modeling, a multi-modal adaptive pixel representation model based on a codebook approach is used. This involves using two dynamically growing vectors of codewords called codebooks for modeling each pixel in the image. For the *RGB* input a codeword is represented by: the average pixel *RGB* value and by the luminance range I_{low} and I_{hi} allowed for this particular codeword. An incoming pixel can be considered to belong to background, if an incoming pixel is within the luminance range and the dot product of $_{pR}GB$ and *RGB* of the codeword is less than a predefined threshold. The codeword for the thermal monochromatic input is represented by: intensity range T_{low} and T_{hi} occurring at the pixel location. Unlike for the color code words the matching of the incoming pixels temperature p_T is done by comparing the ratios of p_T / T_{low} and p_T / T_{hi} to the empirically set thresholds. This method of background modeling automatically adjusts the temperature ranges in thermal codewords to reflect the changing environment. The details of the foreground mask computation are described in [19].

After successful detection of pedestrian, the tracking problem in the Bayesian framework is formulated as the maximization of posteriori probability of the Markov chain state. Bayesian inference process is efficiently implemented by modeling the system as a Markov chain $M = \{x, z, x_0\}$ with a variant of Metropolis-Hastings particle filtering algorithm [18]. As described in [19], in this model, the state of the system at each frame is an aggregate of the state of each body $x_t = \{b_1, \ldots, b_n\}$. Each body, in order, is parametrically characterized as $b_i = \{x, y, h, w, c\}$, where x, y are coordinates of the body on the floor map, h, w its width and height measured in centimeters and c is a 2D color histogram, represented as 32 by 32 bins in hue-saturation space [19]. The body is modeled by the ellipsoid with the axes h and w. An additional implicit variable of the model state is the number of tracked bodies n. The goal of the tracking system is then to find the candidate state x' (a set of bodies along with their parameters) which, given the last known state x, will best fit the current observation z. Therefore, at each frame we aim to maximize the posterior probability:

$$P(x' \mid z, x) = P(z \mid x').P(x' \mid x) \qquad (1)$$

According to Bayes rule and given (1) we formulate our goal as finding:

$$x' = \arg\max_{x'}(P(z \mid x').P(x' \mid x)) \quad (2)$$

The right hand side of equation (2) is comprised of the observation likelihood and the state prior probability. They are computed as joint likelihoods for all bodies present in the scene [19].

Next, the particle filtering approach - a nondeterministic multivariate optimization method is used to solve the optimization problem. The particle filter based method renders the system robust if the joint distribution is not known explicitly, allowing different (Metropolis-Hastings) sampling

$$\alpha(x, x') = \min\left(1, P(x') \Big/ P(x_t).\frac{m_t(x \mid x')}{m_t(x' \mid x)}\right) \qquad (4)$$

Where x' is the candidate state, $P(x)$ is the stationary distribution of the Markov chain, m_t is the proposal distribution. In equation (4), the first part is the likelihood ratio between the proposed sample x' and the previous sample x_t. The second part is the ratio of the proposal density in both directions (1 if the proposal density is symmetric). Further details of the particle filter approach are described in [19].

4. Experimental Results

The evaluation of two approaches described in Sections 3 and 4 was carried out on the OTCBVS benchmark- OSU thermal pedestrian database [1,2,3,19] (acquired by Raytheon 300D thermal sensor and available at http://www.cse.ohio-state.edu/otcbvs-bench/). OSU thermal database has 10 test sequences. Each sequence contains 18–73 frames that are taken within one minute. This database reasonably covers a variety of environmental conditions such as rainy, cloudy and sunny days. In OSU database, the camera is kept still all the time, and the camera–pedestrian distance is far. The set contains short outdoor pedestrian sequences in two locations. Each scene is filmed both with a RGB and thermal camera at the identical resolution, providing thus a pixel to pixel correspondence between two types of sensors. The detection results for different sequences in the colour-thermal dataset for the OSU dataset for the two approaches are shown in Table I and II. We used the same terminology used in [20] for facilitating the evaluation of two approaches. #TP is number of true positives, #FP is number of false positives, PPV is the positive predictive value computed as (PPV = 1 - #FP/#People), and Sensitivity as #TP/#People.

Table I: %TP and %FP for Shape and Appearance Transformation method (SATM) vs. Particle Filter method(PFM)

Sequence	#TP% SATM	#TP% PFM	#FP% SATM	#FP% PFM
1	95.5%	96.8%	4.5%	3.2%
2	94.2%	97.9%	5.8%	2.1%
3	98.2%	99.16%	1.8%	0.84%
4	98.65%	99.5%	1.35%	0.5%
5	95.87%	99.8%	4.13%	0.2%
6	97.8%	99.4%	2.2%	0.3%
ALL SEQ,	94.5%	99.49%	5.5	0.51%

As can be seen in Table and Table 2, both approaches perform satisfactorily, however, the particle filtering based approach performs slightly better than the shape and appearance based approach for this test sequence

Table II: *Sensitivity(ST) and Positive Predicted Value (PPV) for SATM vs. PFM*

Sequence	ST SATM	ST PFM	PPV SATM	PPV PFM
1	0.951	0.968	0.55	0.68%
2	0.942	0.979	0.42	0.58
3	0.982	0.992	0.82	0.16
4	0.986	0.995	0.865	0.5
5	0.958	0.998	0.587	0.8
6	0.978	0.994	0.78	0.7
ALL SEQ,	0.945	0.995	0,45	0.49

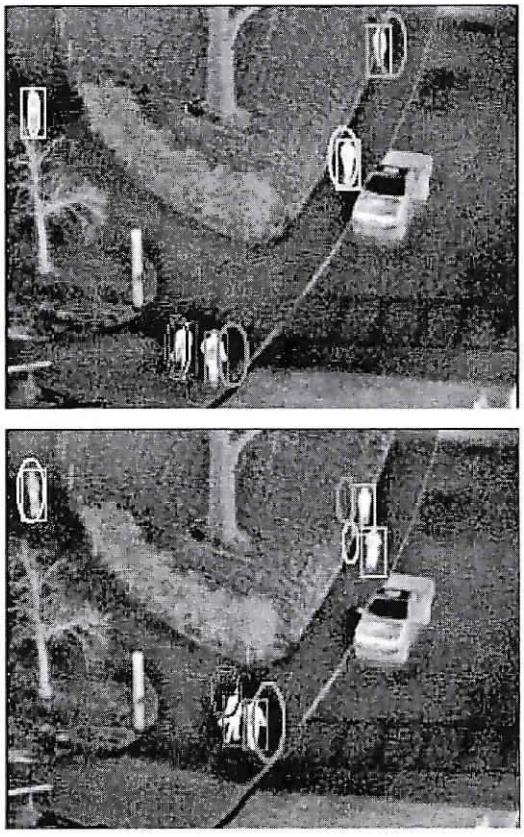

Fig.2(a)I: Tracking performance for Shape and Appearance Transformation method (SATM) vs. Particle Filter method(PFM) for frame 13 and 14.

Fig. 2a to 2c shows the tracking result for frames numbered. 13–18 of sequence #5 in OSU database. Each pedestrian is marked by a different colour.

Further, the results for Particle filtering method is shown with an oval marker and shape and appearance based approach with a square marker. It can be observed that pedestrians tracking performance is satisfactory for frames with no overlapping with other objects or when two pedestrians are located far from each other Also, thee tracking works correctly only if the pedestrian is detected successfully, It will not work if some pedestrian is missed at the detection stage.

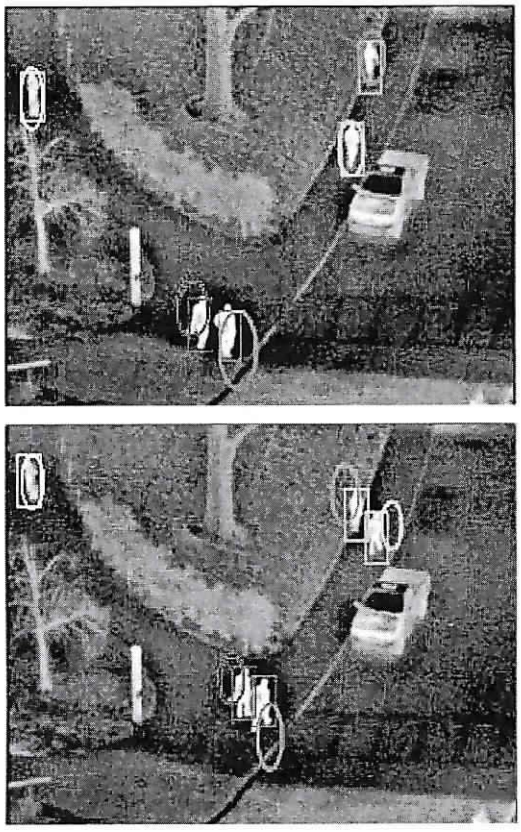

Fig.2(b): Tracking performance for Shape and Appearance Transformation method (SATM) vs. Particle Filter method(PFM) for frame 15 and 16.

5. Conclusions

In this paper we examined a particle filter based approach for pedestrian detection and tracking based on multisensory fusion of colour and thermal datasets of [1,2,19] and compared it with a shape and appearance based approach [20]. Further work is in progress in development of multistage pedestrian detection and tracking approach based on fusion of both shape and appearance based approach with particle filter based Bayesian approach for more complex datasets in the database.

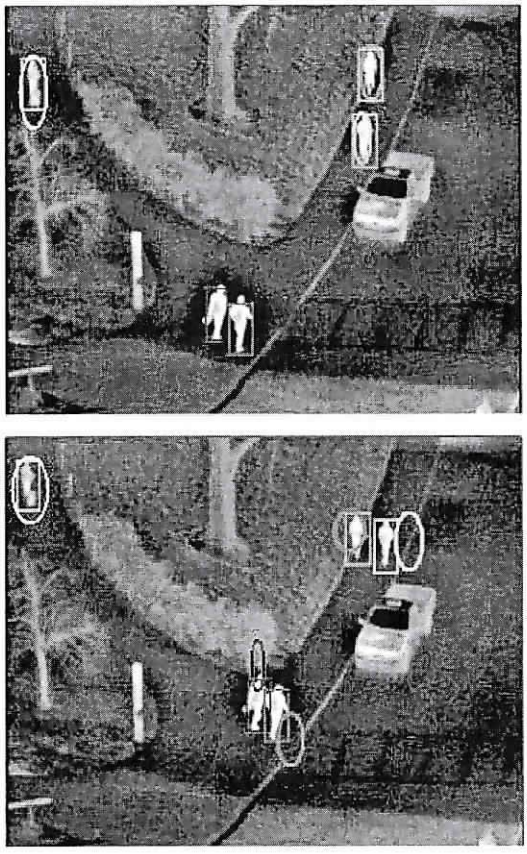

Fig.2(c): Tracking performance for Shape and Appearance Transformation method (SATM) vs. Particle Filter method(PFM) for frame 17 and 18.

6. References

[1] C. Dai, Y. Zheng, and X. Li. "Layered representation for pedestrian detection and tracking in infrared imagery". In IEEE CVPR WS on OTCBVS, 2005.

[2] J. Davis and V. Sharma. "Fusion-based background subtraction using contour saliency". In IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, IEEE OTCBVS WS Series Bench, 2005.

[3] J. Davis and V. Sharma, "Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery," Computer Vision and Image Understanding, Vol 106, No. 2-3, 2007, pp. 162-182.

[4] R. T. Collins, A. J. Lipton, H. Fujiyoshi, T. Kanade, "Algorithms for cooperative multisensory surveillance", Proc. IEEE, Vol. 89, No. 10, pp. 1456-1477, Oct. 2001;

[5] J. M. Ferryman, S. J. Maybank, A. Worrall, "Visual surveillance for moving vehicles", Intl. J. Compu. Vis., Vol. 37, No. 2, pp. 187-197, Oct. 1998;

[6] S. Nadimi, B. Bhanu, "Physics-based models of color and IR video for sensor fusion", Proc. IEEE Multisensor Fusion and Integration for Intelligent systems, MFI'03, pp. 161-166, July 2003;

[7] I. A. Essa, "Ubiquitous sensing for smart and aware environments: technology towards the building of an aware home", IEEE Personal Communications, pp. 47-49, October 2000;

[8] J. W. Fisher III, T. Darrell, "Signal level fusion for multimodal perceptual user interface", Proc. ACM PUI 2001, Orlando, FL USA;

[9] A. Garg, V. Pavlovic, J. Rehg, "Boosted learning in dynamic Bayesian networks for multimodal speaker detection", Proc. IEEE, Vol. 91, No. 9, pp. 1355-1369, Sep. 2003;

[10] D. G. Stork, G. Wolff, E. Levine, "Neural network lipreading system for improved speech recognition", Proc. Intl. Conf. on Neural Networks, IJCNN'92, Vol. 2, pp. 289-295, 1992;

[11] R. Cutler, L. Davis, "Look who's talking: Speaker detection using video and audio correlation", Proc. IEEE Intl. Conf. Multimedia and Expo. ICME'00, pp. 1589-1592, 2000;

[12] T. Ikeda, H. Ishiguro, M. Asada, "Attention to clapping - a direct method for detecting sound source from video and audio", Proc. of IEEE Intl. Conf. on Multisensor Fusion and Integration for Intelligent Systems, MFI'03, pp. 26- 268, 2003;

[13] M. Isard and J. MacCormick. Bramble: "A bayesian multipleblob tracker". In International Conference on Computer Vision, 2001.

[14] T. Zhao and R. Nevatia. "Tracking multiple humans in crowded environment". In International Conference on Computer Vision and Pattern Recognition, 2004.

[15] C. Kemp and T. Drummond. "Multi-modal tracking using texture changes". In British Machine Vision Conference, 2004.

[16] C. Sminchisescu and B. Triggs. "Kinematic jump processes for monocular 3d human tracking". In International Conference on Computer Vision and Pattern Recognition, 2003.

[17] A. Leykin and M. Tuceryan. "A vision system for automated customer tracking for marketing analysis: Low level feature extraction". In Human Activity Recognition and Modelling Workshop, 2005.

[18] A. Elgammal and L. Davis. "Probabilistic framework for segmenting people under occlusion". In International Conference on Computer Vision, 2001.

[19] Alex Leykin, Riad Hammoud, "Robust Multi-Pedestrian Tracking in Thermal-Visible Surveillance Videos", DOI Bookmark:http://doi.ieeecomputersociety.org/10.1109/CVPRW.2006.175

[20] C. Dai, Y. Zheng, X. Li "Pedestrian detection and tracking in infrared imagery using shape and appearance", Elsevier Computer Vision and Image Understanding 106 (2007) 288–299.