Tracking multi-person robust to illumination changes and occlusions

Kyoung-Mi Lee **Youn-Mi Lee**

Department of Computer Science
Duksung Women's University
Seoul 132-714, Korea
kmlee, blanchia@duksung.ac.kr

Abstract

The majority of conventional video tracking surveillance systems assumes a likeness to a person's appearance for some time, and individuality of the person during tracking. However, illumination changes by weather or lighting and occlusions by overlapping people make tracking difficult. To address this situation, we use an adaptive noise, background, and human body model updated statistically frame-by-frame, and correctly construct a person with body parts. Each incoming frame image is corrected to remove shadows by illumination changes and to make a noise image. The corrected image is subtracted by a background model to detect persons. The detected persons are formed by a human body model. The formed person is labeled and recorded in a person's list, which stores the individual's human body model details. Such recorded information can be used to identify tracked persons. The results of this experiment are demonstrated in several indoor situations.

Key words: Adaptive noise model, color-based blob, relation-based person, multiple persons tracking, adaptive body model

1. Introduction

Tracking people based on images using a video camera plays an important role in surveillance systems [2,3]. Many approaches have been proposed to track the human body [1,5]. A popular approach of tracking persons using a fixed camera consists of three steps: background subtraction, blob formation, and blob-based person tracking. To look for regions of change in a scene, a tracking system builds a background model as a reference image and subtracts the monitored scene from the reference frame. However, image frames are easily corrupted by illumination changes and background subtraction is extremely sensitive to dynamic scene changes due to lighting and extraneous events. Blob formation involves grouping homogeneous pixels based on position and color. However, it is still difficult to form blobs of individual body parts using such spatial and visual information since color and lighting variations causes tremendous problems for automatic blob formation algorithms. The tracking human body parts is performed by relating body parts in consecutive frames [7,8]. Such correspondence-based approaches work well for cases with no occlusion, but are unable to decide upon a person identity if a human body is occluded.

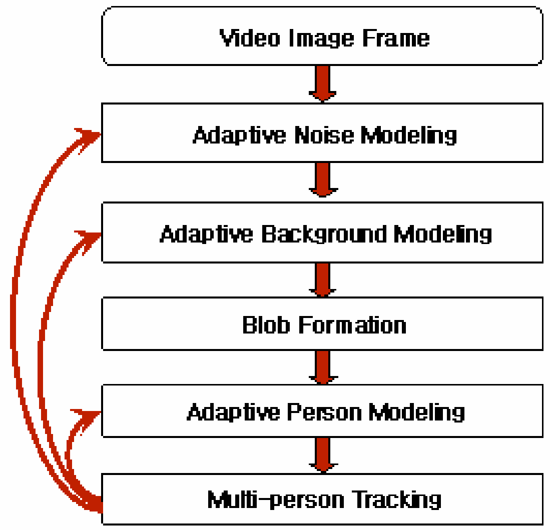


Fig. 1 Flowchart of the proposed approach for tracing multi-person

In this paper, we propose a framework to track multiple persons in a fixed camera situation with illumination changes and occlusion. Fig.1 shows the flowchart of the proposed tracking approach. During tracking, the proposed approach gathers information on noise and background with lighting variations and persons with occlusion, and adaptively updates a background model and person models.

2. Adaptive Noise Modeling

Video image frames taken by a camera have variation in illumination conditions caused by lighting, time of day, and so on. Since noises by such conditions make tracking difficult, the noise should be removed from the image frame.

To separate noises from images, an intrinsic image can be used to get a noise image by subtracting from the image frame. While adding the noise image to the image frame corrects well background, it is not sufficient to correct non-background objects. Recently, Matsushita *et. al* proposed a method for time-dependent intrinsic image estimation [9]. In this paper, we update a noise image frame-by-frame to estimate a time-varying intrinsic image. We first initialize a noise image by subtracting the first image frame from the intrinsic image and then update the noise image frame-by-frame. If a pixel is similar with a noise pixel, the pixel is updated. For visualization, Fig. 2 shows result images after background subtraction (Sec. 3). Fig. 2(b) is a foreground image with illumination correction.

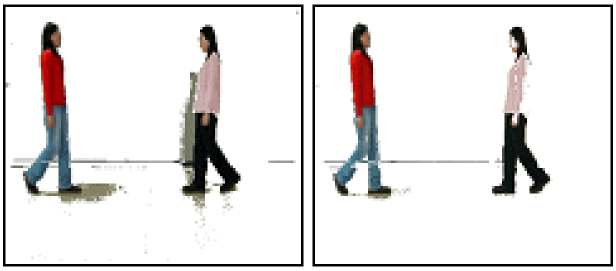

(a) (b)

Fig. 2 Illumination correction: (a) Foreground image with shadow, and (b) corrected image using adaptive noise modeling

3. Adaptive Background Modeling

To detect moving persons in video streams, background subtraction provides the most complete feature data, but it is unfortunately extremely sensitive to dynamic scene changes due to lighting and other extraneous events. For background subtraction, the current image is compared to a reference image to detect changed pixels. During tracking, the reference image should be compensated according to the lighting conditions present in the scene. Thus, the reference image has to adapt to slow changes such as illumination changes by updating the background model. In this paper, we build the adaptive background model statistically, using the mean (μ_B) and standard deviation (σ_B) of the background.

At the pixel level, let I^t represent the intensity value in the *t*-th frame. At time *t*, a change in pixel intensity is computed using Mahalanobis distance δ^n:

$$\delta^t = \frac{\left|I^t - \mu_B^t\right|}{\sigma_B^t} \tag{1}$$

where μ_B^t and σ_B^t are the mean and standard deviations of the background at time *t*, respectively. μ_B^0 is initially set to the first image, $\mu_B^0 = I^0$, and σ_B^0 is initialized by **0**.

Whenever a new frame I^t arrives, a pixel in the frame is tested by Eq. (1) to classify background or foreground (moving persons). If a pixel difference is significantly larger than a predefined threshold at time *t*, the adaptive background model (μ_B^t and σ_B^t) is updated as follows [6]:

$$\begin{aligned} \mu_B^t &= \alpha^{t-1}\mu_B^{t-1} + \left(1-\alpha^{t-1}\right)I^t, \quad \text{and} \\ \sigma_B^t &= \sqrt{\alpha^{t-1}W + \left(1-\alpha^{t-1}\right)\left\{\mu_B^t - I^t\right\}^2}, \end{aligned} \tag{2}$$

where $W = \left\{\sigma_B^{t-1}\right\}^2 + \left\{\mu_B^t - \mu_B^{t-1}\right\}^2 \cdot \alpha^{t-1} = \frac{N_B^{t-1}}{N_B^{t-1}+1}$ where N_B^{t-1} means the number of frames participating in the background model to time *t*-1.

4. Person Model Initialization

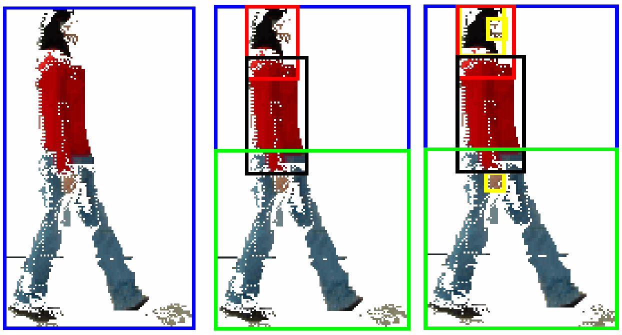
Before tracking persons, they should be initialized when they start to appear in the video. To group segmented foreground pixels into a blob and to locate the blob on a body part, we use a connected-component algorithm which calculates differences between intensities between a pixel and its neighborhoods. However, it is difficult to form perfect individual body parts using such a color-based grouping algorithm since color and lighting variations causes tremendous problems for automatic blob formation algorithms. Therefore, small blobs are merged into large blobs and neighboring blobs that share similar colors are further merged together to overcome over-segmentation generated by initial grouping, which is largely due to considerable illumination changes across the surfaces of coherent blobs. Each blob B_i contains information such as an area, a central position, a bounding box, and a boundary to form a human body. Then, some blobs are removed according to criteria, such as, too small, too long, or too heterogeneous incompact blobs.

As a person can be defined as a subset of blobs, which correspond to human body parts, blobs in a frame should be assigned to corresponding individuals to facilitate multiple individual tracking. The person formation algorithm first computes a blob adjacency graph and a minimum distance between the associated bounding boxes of each blob. A distance between the vertices of the bounding boxes is computed much faster than adjacency in the pixel domain [4]. Let P_0 be a subset of blobs B_i. The set of potential person areas is built iteratively, starting from the P_0 set and its adjacent graph. The set P_1 is obtained by merging compatible adjacent blobs of P_0. Then, each new set of merged blob P_k is obtained by merging the set of merged blob $P_{k\text{-}1}$ with the original set P_0. Finally, a candidate person CP_n contains the built sets of merged blob P_k, i.e., $CP_n = \bigcup_{k=0}^{K} P_k$, n=1…N where N is the number of persons.

To match blobs to body parts, we use a hierarchical person model (see Fig. 3). First, we assumed that all individuals in the video are upright or in a slightly skewed standing position. The high level of the model contains a whole person model and its information (Fig. 3(a)). Depending on the relative position in the model, each blob is assigned to one of three body parts in the middle level: the head, the upper body, and the lower body (Fig. 3(b)). If a body part in the middle level contains two more blobs, these blobs are tested for skin similarity to classify the blobs as skin or non-skin area at the low level (Fig. 3(c)). The person model is defined by three body parts and their geometrical relations in the middle level, and as blobs and their geometrical and color relations in the low level as follows:

$$CP_n = \left(R_n^0, \left\{C_n^1, R_n^1\right\}, \left\{C_n^2, R_n^2\right\}, \left\{C_n^3, R_n^3\right\}\right) \qquad (3)$$

where R^0 means a relation among three parts. C_n^j and R_n^j mean a set of blobs and their relationships of the j-th body part of CP_n, respectively.



(a) (b) (c)

Fig. 3 A hierarchical person model: (a) Bounding region in the high level after background subtraction, (b) three body parts in the middle level, and (c) skin blobs in the low level

5. Adaptive Person Modeling

Since a person is tracked using the person model (Eq. (3)), person model information is stored to track multiple-persons. Even though the total motion of a person is relatively small between frames, large changes in the color-based model can cause simple tracking to fail. To resolve this sensitivity of color-model tracking, we compare the current blobs to a reference person model. Like an adaptive background model (Sec. 3), the reference person model should be compensated according to occlusion as well illumination changes.

Let a person CP_n^{t-1} represented by an average (μ_n^{t-1}) and a deviation (σ_n^{t-1}), which are computed up to time t-1 and new blobs B_i^t and their relations Br_i^t are formed in frame t. The minimum difference between the person model CP_n^{t-1} (Eq. (3)) and the new blobs B_i^t is computed as follows:

$$d_n^t = \min_{j=1\cdots3}\left(\frac{\left\| B_i^t - \mu_n^{t-1,C^j}\right\|_p}{\sigma_n^{C^j}}\right) + \min_{j=0\cdots4}\left(\frac{\left\| Br_i^t - \mu_n^{t-1,R^j}\right\|_p}{\sigma_n^{R^j}}\right) \qquad (4)$$

where μ_n^{t-1,C^j} and μ_n^{t-1,R^j} mean a set of averages of blobs and relations in the j-th body part at time t-1, respectively. σ_n^{t-1,C^j} and σ_n^{t-1,R^j} a set of deviations of blobs and relations, respectively. If the minimum distance is less than a predefined threshold, the proposed modeling algorithm adds blobs B_i^t and relations Br_i^t to corresponding adaptive person model (μ_n^{t-1,C^j} and μ_n^{t-1,R^j}) and updates the adaptive model by recalculating their center and uncertainties [6]. Here, similarity thresholds are set empirically and can be adjusted by a user.

6. Model-based Multi-person Tracking

Tracking people poses several difficulties, since the human body is a non-rigid form. After forming blobs, a blob-based person tracking maps blobs from the previous frame to the current frame, by computing the distance between blobs in consecutive frames. However, such a blob-based approach for tracking multiple persons may cause problems due to the different number of blobs in each frame: blobs can be split, merged, or even disappear or be newly created. To overcome this situation, many-to-many blob mapping can be applied [8]. While these authors avoided situations where blobs at time t-1 are associated to a blob at time n, or vice versa, they adopted a variant of the multi-agent tracking framework to associate multiple blobs simultaneously.

In this paper, we assume that persons CP_n^{t-1} have already been tracked up to frame t-1 and new blobs B_i^t are formed in frame t. Multi-persons are then tracked as follows:

Case 1: If B_i^t is included in CP_n^{t-1}, the corresponding blob in CP_n^{t-1} is tracked to B_i^t.

Case 2: If a blob in CP_n^{t-1} is separated into several blobs in frame t, the blob in CP_n^{t-1} is tracked to one blob in frame t and other blobs at time t are appended into CP_n^{t-1}.

Case 3: If several blobs in CP_n^{t-1} are merged into B_i^t, one blob in CP_n^{t-1} is tracked to B_i^t and other blobs are removed from CP_n^{t-1}.

Case 4: If B_i^t is included in CP_n^{t-1} but the corresponding blob does not exist, B_i^t is added to CP_n^{t-1}.

Case 5: If B_i^t is not included in CP_n^{t-1}, the blob is considered as a newly appearing blob and thus a new person is added to the person list (Sec. 4).

where including a region into a person with a bounding box means the region overlaps over 90% to the person. Corresponding a blob to the adaptive person model is computed using Eq. (4). In addition to simplify the handling of lists of persons and blobs, the proposed approach can keep observe existing persons exiting, new persons entering, and previously monitored persons re-entering the scene. One advantage of the proposed approach is to relieve the burden of correctly blobbing. Even though a blob can be missed by an illumination change, model-based tracking can retain individual identity using other existing blobs. After forming persons (Sec. 5), the number of blobs in a person is flexibly changeable. The proposed approach can handle over-blobbing (Case 2) and under-blobbing (Case 3) problems.

7. Results and Conclusions

The proposed multi-person tracking approach was implemented in JAVA (JMF), and tested in Windows 2000 on a Pentium-IV 1.8 GHz CPU with a memory of 512 MB. For 320×240 frames, videos were recorded using a Sony DCR-PC330 camcoder.

Fig. 4 presents tracking results in indoor environments with occlusion and shadow. Whenever a person appears first in the video, a new person model is built and added to the persons' list (Fig. 4(a)). During tracking the person, the proposed adaptive tracking system updates the corresponding person model at each frame. Even though two persons in a frame are occluded by crossing over or hugging (Fig. 4(b) and (c)), the proposed system can track the persons by keeping the person's information. Fig. 3 contains shadow near to shoes on the left of frames (Fig. 4(a) and (d)), and the person model includes the shadow as a connected component. But, such a noise in the person model does not affect to track and identify persons.

The goal of this paper was to track multi-person in environments of changing illumination and occlusions. By adaptively building noise, background and person models, we can successfully track multi-person even if persons' information is partially lost.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2004-003-D00376).

References

[1] J.K. Aggarwal and Q. Cai, "Human motion analysis: a review", *Computer vision and image understanding*, 73(3):428–440, 1999.

[2] G. Foresti, P. Mahonen, and C. S. Regazzoni, "Multimedia video-based surveillance systems: requirements, issues and solutions", Dordrecht, The Netherlands: Kluwer, 2000.

[3] G. Foresti, P. Mahonen, and C. S. Regazzoni, "Automatic detection and indexing of video-event shots for surveillance applications", *IEEE transactions on Multimedia*, 4(4):459-471, 2002.

[4] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis", *IEEE transactions on Multimedia*, 1(3):264-277, 1999.

[5] D. Gavrila. "The visual analysis of human movement: a survey", *Computer vision and image understanding*, 73(1):82–98, 1999.

[6] K.-M. Lee and W. N. Street, "Model-based detection, segmentation, and classification using on-line shape learning", *Machine vision and applications*, 13(4):222-233, 2003.

[7] W. Niu, L. Jiao, D. Han, and Y.-F. "Wang, Real-time multi-person tracking in video surveillance", *Proceedings of the Pacific Rim multimedia conference*, 1144-1148, 2003.

[8] S. Park and J.K. Aggarwal, "Segmentation and tracking of interacting human body parts under occlusion and shadowing", *Proceedings of InternationalWorkshop on motion and video computing*, 105-111, 2002.

[9] Y. Matsushita, K. Nishino, K. Ileuchi, and S. Sakauchi, "Illumination normalization with time-dependent intrinsic images for video surveillance", *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(10):1336-1347, 2004.

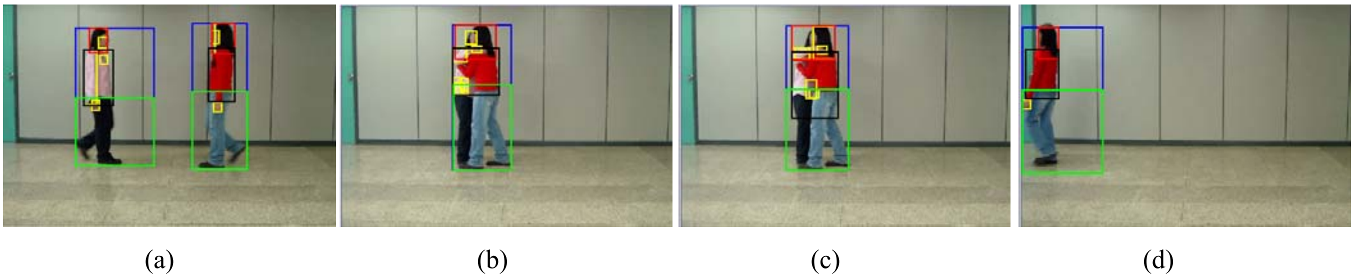


Fig. 4 Multi-person tracking in presence of illumination changes and occlusion: (a) two persons are entering in the scene, (b) and (c) two persons are occluded by hugging, and (d) one person is exiting with shadow.