Computer Vision and Image Understanding **80,** 42–56 (2000)
doi:10.1006/cviu.2000.0870, available online at http://www.idealibrary.com on IDEAL®

Tracking Groups of People

Stephen J. McKenna

Department of Applied Computing, University of Dundee, Dundee DD1 4HN, Scotland

Sumer Jabri and Zoran Duric

Department of Computer Science, George Mason University, Fairfax, Virginia 22030-4444

Azriel Rosenfeld

Center for Automation Research, University of Maryland, College Park, Maryland 20742-3275

and

Harry Wechsler

Department of Computer Science, George Mason University, Fairfax, Virginia 22030-4444

Received November 2, 1999; accepted July 5, 2000

A computer vision system for tracking multiple people in relatively unconstrained environments is described. Tracking is performed at three levels of abstraction: regions, people, and groups. A novel, adaptive background subtraction method that combines color and gradient information is used to cope with shadows and unreliable color cues. People are tracked through mutual occlusions as they form groups and separate from one another. Strong use is made of color information to disambiguate occlusion and to provide qualitative estimates of depth ordering and position during occlusion. Simple interactions with objects can also be detected. The system is tested using both indoor and outdoor sequences. It is robust and should provide a useful mechanism for bootstrapping and reinitialization of tracking using more specific but less robust human models. © 2000 Academic Press

1. INTRODUCTION

Visual surveillance and monitoring of human activity requires people to be tracked as they move through a scene. Such a tracking system can be used to learn models of activity from extended observations. These activity models can then be used to detect unusual or important events [9, 16] and to constrain tracking. Interactions with objects that are often of interest include picking up an object, placing an object in the scene, or passing an object to

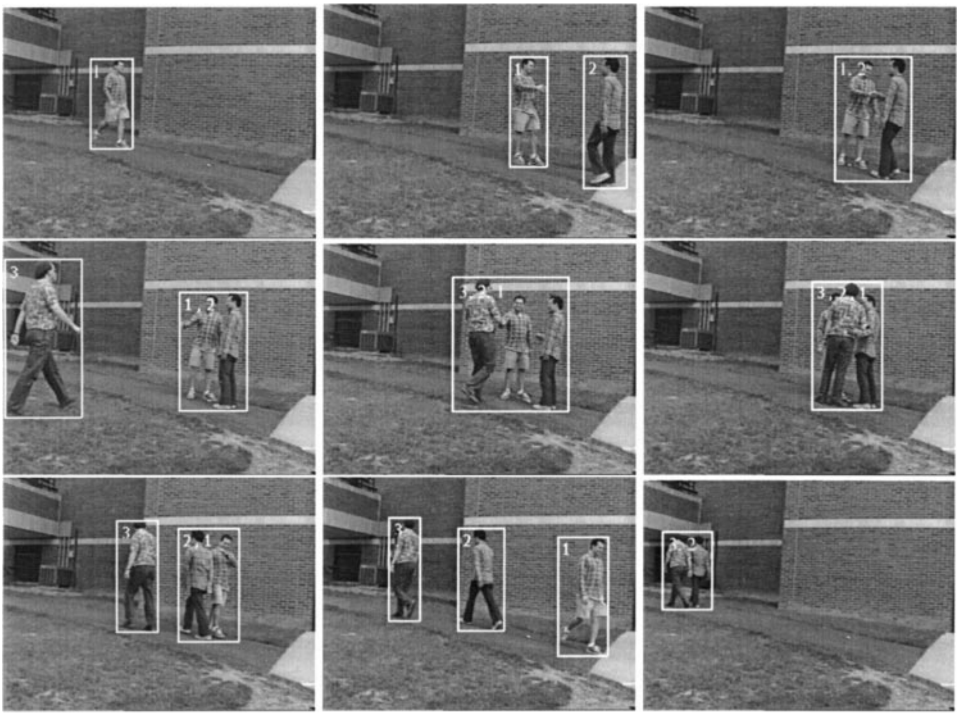

FIG. 1. Frames from a sequence of people being tracked as they form groups.

another person. Tracking people in relatively unconstrained, cluttered environments as they form groups, interact, and part from one another requires robust methods that cope with the varied motions of the humans, occlusions, and changes in illumination. The significant (sometimes complete) occlusions that occur when people move in groups or interact with other people cause considerable difficulty to many tracking schemes. However, a system capable of understanding the activities of interacting people needs to cope with such situations routinely. Figure 1 shows an example scenario.

Robust tracking of multiple people through occlusions requires person models that are sufficiently specific to disambiguate occlusions. However, we would like the models to be general and simple enough to allow robust, real-time tracking. The models must cope with reasonable changes in illumination, large rotations in depth (people turning to face in a new direction, possibly while occluded by other people), and varied clothing. In particular, for outdoor surveillance (especially in cold climates) it is not reasonable to assume that people wear tightly fitting garments. In a large winter coat, for example, the articulated structure of the body may barely be discernible. Nor is it reasonable to assume that garments are uniformly colored. Homogeneous “blobs” of color will often not correspond well to body parts. However, color and texture do suggest decompositions of humans into visual parts. These parts are often somewhat different from those suggested by shape and motion. They depend more on clothing and vary with a person’s mode of dress. The color distributions of items of clothing are typically quite stable under rotation in depth, scaling, and partial occlusion. Furthermore, color models are easily adapted to account for gradual changes in illumination. This paper presents an efficient, color-based tracking system. It can be viewed as complementary to tracking methods that use more specific human models since it provides a reliable platform from which to bootstrap such methods when sufficient image evidence exists to support their use.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related work. In Section 3, a novel background subtraction technique that fuses color and gradient information is described. A general tracking scheme that uses regions, people, and groups as distinct levels of abstraction is outlined in Section 4. Section 5 describes how adaptive color models of appropriate complexity can be learned on-line and used to track people through occlusions. Example sequences in which people are tracked while interacting with one another and with objects are used to illustrate the approach. Some final comments are made in Section 6.

2. RELATED WORK

Systems for tracking people have usually employed some form of background subtraction: Haritaoglu *et al.* used gray-level subtraction [10, 11], Pfinder modeled pixel color variation using multivariate Gaussians [31], and Gaussian mixtures have also been used in a similar manner [23, 29]. The car tracking system of Koller *et al.* [18] used an adaptive background model based on monochromatic images filtered with Gaussian and Gaussian derivative (vertical and horizontal) kernels. Background "subtraction" using these filter outputs yielded results superior to the use of raw gray levels.

Some systems for surveillance and monitoring of wide-area sites have tracked people by essentially assuming that each connected component obtained from background subtraction (and some further processing) corresponds to a moving object [24, 29]. Trackers based on 2D active shape models have been used but can only cope with moderate levels of occlusion [2, 16]. Color blobs were used in Pfinder to build a person model in a controlled indoor environment [31]. Intille *et al.*'s closed-world tracking was used (for example) to track players on an American football field [13]. McKenna and Gong used a combination of motion, skin color, and face detection to track people and their faces [22]. Lipton *et al.* combined temporal differencing with template matching to track people and cars in wide-area scenes [20]. Darrell *et al.* combined depth from stereo with skin color and face detection [7]. Bregler proposes an ambitious probabilistic framework for tracking at multiple levels of abstraction [3]. Many approaches have been proposed for tracking the human body. The interested reader is referred to reviews of motion understanding approaches [5], hand gesture recognition [25], visual analysis of human motion [8], and nonrigid motion [1].

The Hydra system [11] developed at the University of Maryland is perhaps the most similar in its aims to the system described here. It is essentially an extension of the W4 system [10]. Hydra attempts to detect the heads of people in groups and track them through occlusions. It uses silhouette-based shape models and temporal texture appearance models. Although effective in many situations, these 2D appearance models will not cope well with large rotations in depth during occlusions. Hydra is therefore quite effective at tracking people through occlusions as they walk past one another but it does not cope well when people leave a group in a different direction from that in which they entered it. Hydra uses monochromatic image data. In contrast, the system presented in this paper makes extensive use of color information to build adaptive models that are efficient to compute as well as being stable under rotations in depth, scaling, and illumination changes.

Some other systems have performed occlusion reasoning by assuming that object motion is constrained to a ground plane [18, 28]. This assumption enabled a depth ordering to be estimated so that it was possible to differentiate occluded from occluding objects. In contrast, the occlusion reasoning described here makes no ground plane assumption and therefore has wider applicability.

3. ADAPTIVE BACKGROUND SUBTRACTION

We assume that the camera is stationary and that the background changes only slowly relative to the motions of the people in the scene. The background model combines pixel RGB and chromaticity values with local image gradients. The expected variance of these features is used to derive confidence measures for fusion. The method consists of three stages and produces a foreground segmentation mask.

3.1. RGB Change Detection

The camera's R, G, and B channels are assumed to have Gaussian noise. Three variance parameters σ^2_{rcam}, σ^2_{gcam}, σ^2_{bcam} are estimated for these channels. Some background pixel can violate the Gaussian assumption because of jitter or small "micromotions" such as leaves moving on a tree or waves on water. These changes occur on a short timescale and so cannot be handled using adaptation. Instead they can be considered to give rise to multimodal, stationary distributions for the pixel's values. While these distributions can be modeled as Gaussian mixtures, for example [23, 29], this is usually not worth the added computational expense since small, isolated regions of jitter or micromotion can be discarded during grouping. Instead, we estimate variance parameters for each pixel and these variance parameters are used for background subtraction only when they are greater than the variance due to camera noise. The stored color background model for a pixel is $[\mu_r, \mu_g, \mu_b, \sigma^2_r, \sigma^2_g, \sigma^2_b]$.

Changes in illumination are assumed to occur slowly relative to object motion. The background model is adapted on-line using simple recursive updates in order to cope with such changes. Adaptation is performed only at image locations that higher level grouping processes label as being clearly within a background region. Recursive estimation of mean and variance can be performed using the following update equations given the latest measurement z_{t+1} at time $t + 1$ [19, 26]:

$$\mu_{t+1} = \alpha\mu_t + (1-\alpha)z_{t+1} \tag{1}$$

$$\sigma^2_{t+1} = \alpha\left(\sigma^2_t + (\mu_{t+1} - \mu_t)^2\right) + (1-\alpha)(z_{t+1} - \mu_{t+1})^2. \tag{2}$$

These updates estimate a nonstationary Gaussian distribution. The mean, μ, and the variance, σ^2, can both be time varying. The constant α is set empirically to control the rate of adaptation ($0 < \alpha < 1$). This depends on the frame rate and the expected rate of change of the scene. The smaller the value of α, the faster the old data are (exponentially) forgotten. The sequences in this paper were processed using $\alpha = 0.9$. Given a new pixel at time $t + 1$ with RGB values $(r_{t+1}, g_{t+1}, b_{t+1})$, each of the means and variances in the background model are updated using Eqs. (1) and (2).

The background model can be used to perform background subtraction as follows. The current pixel $\mathbf{x} = (r, g, b)$ is compared to the model. If $|r - \mu_r| > 3\max(\sigma_r, \sigma_{\text{rcam}})$, or if the similar test for g or b is true, then the pixel is set to foreground. Otherwise it is set to background. This produces a mask that is considered as a region of interest for further processing.

3.2. *Gradient and Chromaticity*

The assumption that illumination changes slowly is violated when the change is due to shadows cast by people moving in the scene. Ideally, we would like our background subtraction method not to label such regions of shadow as foreground. An area cast into shadow often results in a significant change in intensity without much change in chromaticity. This observation has been exploited by previous authors to label as shadow pixels that become darker without significant chromaticity change [12, 31]. Our approach exploits a similar assumption but is somewhat different. As shadows appear and disappear, intensity levels decrease and increase. Therefore, we assume that any significant intensity change without significant chromaticity change could have been caused by shadow. Chromaticity is computed as

$$r_{\mathrm{c}} = r/(r+g+b) \tag{3}$$

$$g_{\mathrm{c}} = g/(r+g+b), \tag{4}$$

and each pixel's chromaticity is modeled using means and variances $\mu_{r\mathrm{c}}, \mu_{g\mathrm{c}}, \sigma^2_{r\mathrm{c}}, \sigma^2_{g\mathrm{c}}$. Adaptive background subtraction is performed as before but using chromaticity values instead of RGB values.

Often there will be no difference in chromaticity between foreground and background (e.g., a dark green coat moves in front of grass, or black trousers cross a gray concrete path). In such cases, we cannot reliably tell based on zeroth-order, pixel-level color information whether the pixel has changed due to shadow. However, the use of first-order image gradient information enables us to cope with such cases more effectively.

Gradients are estimated using the Sobel masks in the x and y directions. Each background pixel's gradient is modeled using gradient means (μ_{xr}, μ_{yr}), (μ_{xg}, μ_{yg}), (μ_{xb}, μ_{yb}) and magnitude variances $\sigma^2_{\mathrm{gr}}, \sigma^2_{\mathrm{gg}}, \sigma^2_{\mathrm{gb}}$. Additionally, we compute average variances $\bar{\sigma}^2_{\mathrm{gr}}, \bar{\sigma}^2_{\mathrm{gg}}, \bar{\sigma}^2_{\mathrm{gb}}$ over the entire image area. Adaptive background subtraction is performed as follows. Given a new pixel value $\mathbf{x} = (r, g, b)$, its spatial gradients (r_x, r_y), (g_x, g_y), (b_x, b_y) are estimated using the Sobel operator. If $\sqrt{(r_x - \mu_{xr})^2 + (r_y - \mu_{yr})^2} > 3\max\{\sigma_{\mathrm{gr}}, \bar{\sigma}_{\mathrm{gr}}\}$, or if the similar test for (g_x, g_y) or (b_x, b_y) is true, then the pixel is set to foreground. Otherwise it is set to background.

A pixel is flagged as foreground if either chromaticity or gradient information supports that classification. A detailed description of a similar background subtraction method is given in [14]. This approach helps to eliminate some types of shadows. Shadows with hard edges will still be detected as foreground. However, these tend to be near the person and so cause only small errors during grouping. The long shadows that would cause the greatest problems for grouping tend to have significant penumbras and these soft edges are not detected.

Figure 2 shows an example of background subtraction from a sequence of a person walking and casting a shadow. The center image shows the connected components detected using an adaptive RGB background model. Much of the shadow is labeled as foreground. The rightmost image shows the result when gradient and chromaticity information are combined. Although much of the person's clothing is almost gray (with low chromatic content), the connected component detected is a reasonably good segmentation. Only a very small area of shadow near the person's feet is detected.

The images resulting from background subtraction are filtered using a 3 × 3 median filter, and a connected-component-labeling algorithm is then applied [15]. Any connected component whose area is less than a threshold, T_{cc}, is discarded as being "noise." A contour

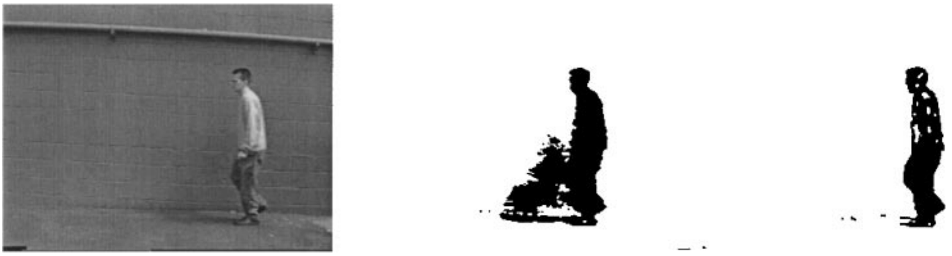

FIG. 2. Example of background subtraction. (Left) Color image from sequence. (Center) Connected components using RGB background model. (Right) Connected components using combined chromaticity and gradient background model.

collecting algorithm can then be used to delineate foreground objects and thus perform a background–foreground segmentation.

A foreground connected component, C, is likely to contain holes. Some of these holes will be due to erroneous segmentation in regions of low chromatic content and low texture. However, holes can also correspond to true background regions when the body is in certain postures, e.g., a frontal view of a person with a hand on a hip. Note that it is possible to detect such holes provided that the background region to which they correspond is not cast into shadow. This can be done by referring to the mask, M_{rgb}, produced by the RGB subtraction described in Section 3.1. Each "background" pixel which is part of a hole contained in a component C is set to foreground if the mask M_{rgb} indicates that it is foreground.

Figure 3 shows the resulting segmentation for the sequence shown in Fig. 1. In all the experiments reported in this paper, adaptive background subtraction was performed using

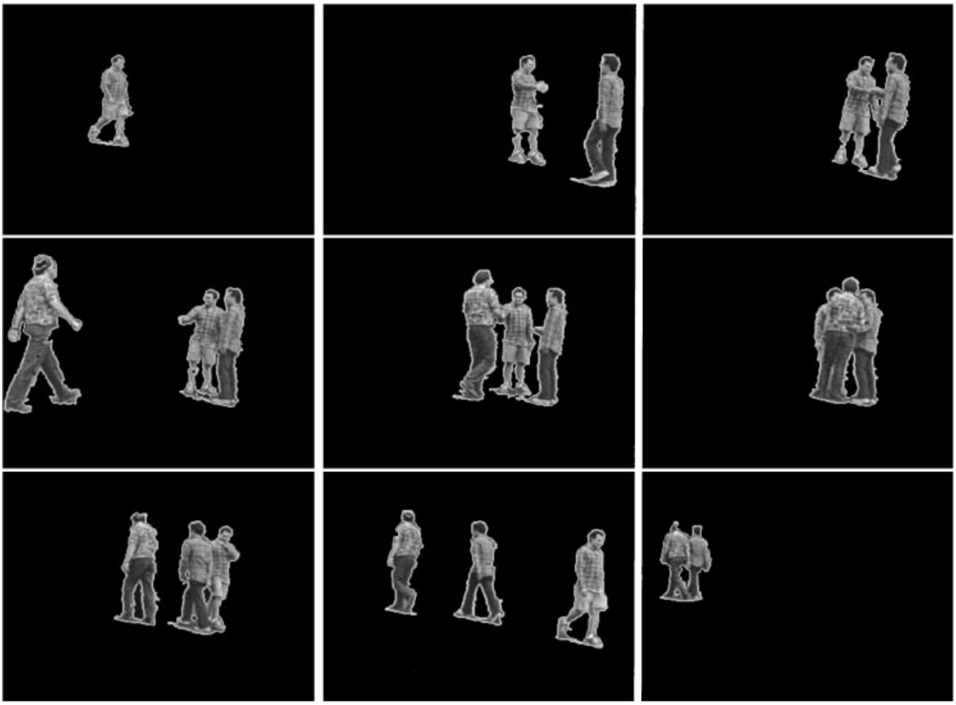

FIG. 3. Segmented frames from a sequence of people being tracked as they form groups.

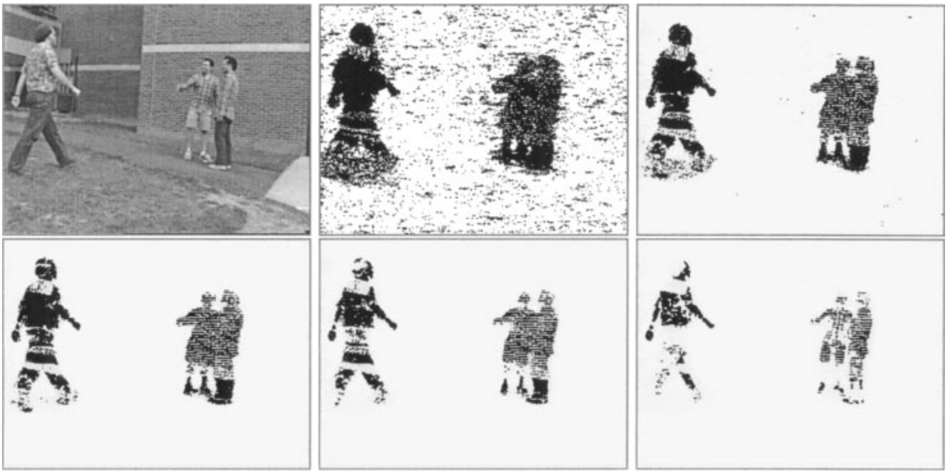

FIG. 4. The effect of varying the threshold on color-based subtraction. From top left: original image (same as fourth frame in Fig. 1) and results of thresholding at 1, 2, 3, 5, and 10 standard deviations.

thresholds at 3σ. In order to help motivate this choice, Fig. 4 shows the effect of varying these thresholds. Thresholding at σ resulted in too much background being classified as foreground. Thresholding at or above 5σ resulted in too much foreground being classified as background. Thresholding at or around 3σ provides an acceptable trade-off. If data are normally distributed, 99.7% of the data are within 3σ of the mean. Furthermore, Chebyshev's theorem guarantees that 88.9% of the data are within 3σ of the mean regardless of the distribution.

4. TRACKING THE FOREGROUND

Robust tracking requires multiple levels of representation. Sophisticated models should only be employed when support is available for them. In fact, many of the articulated geometric models used for human tracking have had to be initialized by hand in the first frame of the sequence, e.g., [4, 6]. A robust, integrated system needs less specific models for tracking; these can be used for initialization and reinitialization of more complex models. The system described here is not concerned with fitting models of articulated body structure or accurate labeling of body parts. However, it complements such approaches. We perform tracking at three levels of abstraction:

Regions. Regions are connected components that have been tracked for at least T_{fr} frames. Each region has a bounding box, a support map (mask), a timestamp, and a tracking status.

People. A person consists of one or more regions grouped together. Each person has an appearance model based on color.

Groups. A group consists of one or more people grouped together. If two people share a region, they are considered to be in the same group.

A temporally consistent list of tracked regions is maintained during tracking. Temporal matching is performed based on the support map and bounding box. In practice, simply matching regions with overlapping bounding boxes was found to be effective. In particular, prediction was not needed since the visual motions of regions were always small relative to their spatial extents.

In each frame, a new region tracker is initialized for each novel region, if any. Regions with no match are deleted. Regions can split and merge. When a region splits, all the resulting regions inherit their parent's timestamp and status. When regions merge, the timestamp and status are inherited from the oldest parent region. Once a region is tracked for three frames, it is considered to be reliable and is subsequently considered for inclusion in the class of people.

It is oversimplistic to assume that regions correspond to people. A person will often split into multiple regions despite the use of good background subtraction techniques. This is the case even if morphological operations such as opening and closing are used. Therefore, a person is initialized when one or more regions that currently belong to no person satisfy a set of rules. In order to form a person, the regions must be in close proximity, their projections onto the x axis must overlap, and they must have a total area larger than a threshold, T_{person}. Further rules based on aspect ratio and skin color can be added. Once a person is being tracked, any region that overlaps its bounding box (or alternatively its support map) is matched to the person. A person is considered to constitute a group of one.

A group consists of one or more people and therefore one or more regions. Groups can split and merge. When a region is matched to more than one group, those groups are merged to form a new group. When the regions in a group do not have sufficient proximity or do not overlap on the x axis, that group is split up. A split usually results in a large group containing N people dividing into smaller groups that together contain N people. However, regions that contain no people can also split from a group when a person deposits an object, for example.

The thresholds were set to $T_{\text{fr}} = 4$ frames, $T_{\text{cc}} = 30$ pixels, and $T_{\text{person}} = 500$ pixels for the sequences shown in this paper. Figure 5 shows three frames from a sequence in which a man and a woman walk toward one another, greet briefly, and then continue. The tracker successfully tracks them through the sequence. Figure 6 illustrates the processing that is performed on the last three frames of Fig. 5. Although the man is grouped as a single region for most of the sequence, in this particular frame background subtraction splits him into two regions. However, he is correctly tracked as a group consisting of a single person.

Another example sequence that shows three people being tracked as they form groups and split up again was shown in Fig. 1 with a bounding box shown for each tracked group. Figure 7 shows the centers of these boxes in every frame overlaid on the empty scene. The plots in Fig. 8 show how the x and y coordinates of these box centers change over time. The resolution of this sequence was 320 × 240 pixels. In order to obtain a quantitative estimate of tracking accuracy, a bounding box was determined by hand for person 1 for 65 frames after that person had entered the scene. These bounding boxes were treated as ground-truth and were compared to the tracking system's estimates. The mean absolute

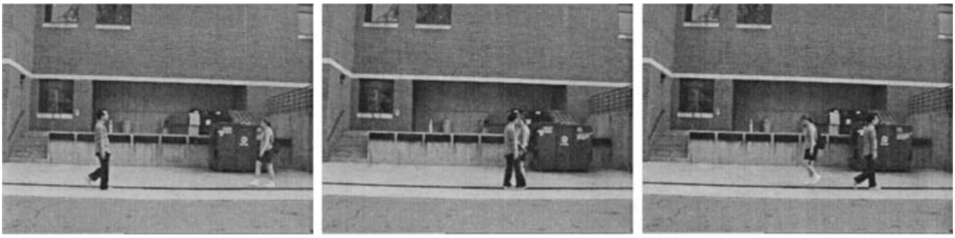

FIG. 5. Three images from a sequence in which two people walk past one another.

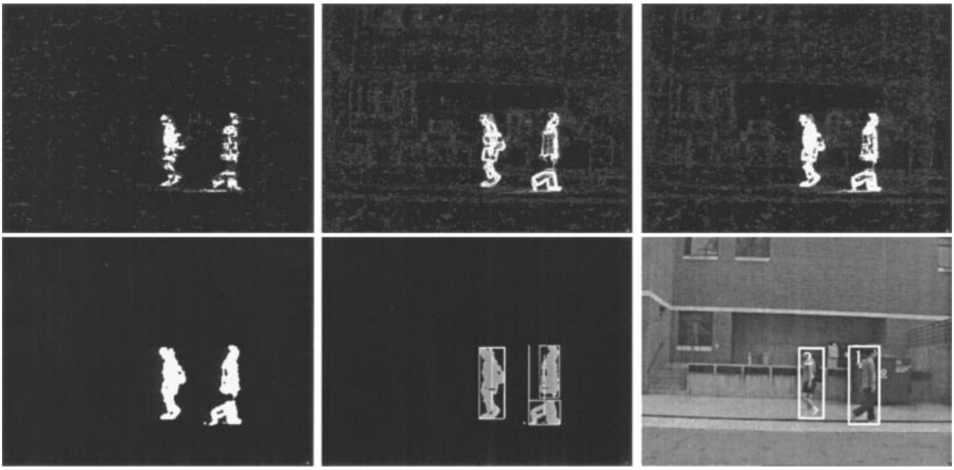

FIG. 6. Processing of the third frame in Fig. 5. From top left: (i) Background subtraction using chromaticity, (ii) background subtraction using gradients, (iii) combined background subtraction, (iv) combined background subtraction after median filtering, (v) bounding boxes for tracked regions and people (person 1 consists of two regions), and (vi) tracked people's bounding boxes.

errors in estimating the x coordinates and y coordinates for the bounding box were 1.6 pixels (standard deviation = 1.4) and 0.7 pixels (standard deviation = 0.7) respectively. This resulted in mean absolute errors for the box center plotted in Fig. 8 of 0.8 pixels for the x coordinate (standard deviation = 0.6) and 0.5 pixels for the y coordinate (standard deviation = 0.4). The maximum absolute errors for any frame were 2.5 and 1.5 pixels, respectively.

The tracking system is also able to detect some interactions with objects. If a person removes an object or deposits an object in the scene, this will give rise to a new region that splits from the person. If this region does not have significant motion and is not part of a person, it is flagged as corresponding to an object that has just been acted upon by the person. Figures 9 and 10 show examples of detected interactions with objects.

5. MODELING THE FOREGROUND

In order to track people consistently as they enter and leave groups, each person's appearance must be modeled. This allows people to be tracked despite the ambiguities that

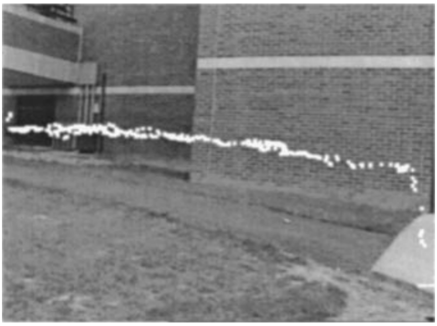

FIG. 7. The centers of the bounding boxes of the groups tracked in Fig. 1 shown overlaid on the empty scene.

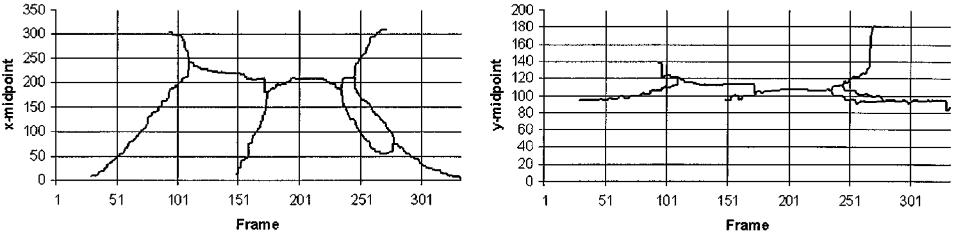

FIG. 8. The coordinates of the bounding box centers of the groups tracked in Fig. 1 plotted over time. (Left) Centers' *x* coordinates. (Right) Centers' *y* coordinates.

arise as a result of occlusion and grouping of people. Therefore, a color model is built and adapted for each person being tracked. Since people cannot be reliably segmented while grouped with others, these person models are adapted only while a person is alone, i.e., in a group of size one.

Color distributions have been effectively modeled for tracking using both color histograms [21] and Gaussian mixture models [23, 27], and these models give similar results. Both histograms and Gaussian mixtures can be updated adaptively. Gaussian mixtures can be estimated from color data using an expectation-maximization algorithm. When adaptation is not needed, a mixture can be used to generate a histogram for fast computation of probabilities. When the number of color samples is small and the number of meaningfully discriminable colors is large (e.g., true 24-bit pixels acquired using a high-quality camera), Gaussian mixtures are more appropriate. Conversely, histograms are appropriate with larger data sets in a coarsely quantized color space. For example, histograms are slightly more effective than mixture models for modeling skin color using very large numbers of images taken from the World Wide Web [17]. We have used both histograms and mixture models effectively.

A histogram $H_i(\mathbf{x})$ simply counts the number of occurrences of $\mathbf{x} = (r, g, b)$ within the mask for person i. It provides a look-up table from which a discrete probability distribution is obtained as

$$P(\mathbf{x} \mid i) = \frac{H_i(\mathbf{x})}{A_i}, \tag{5}$$

where A_i is the area of the person mask in pixels. In each frame, such a model can be updated either cumulatively to model a stationary distribution or, more appropriately, adaptively to model a nonstationary distribution. Histogram models are adaptively updated by storing the

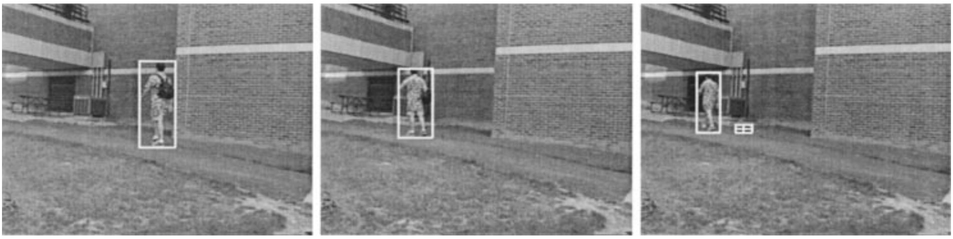

FIG. 9. A person deposits an object.

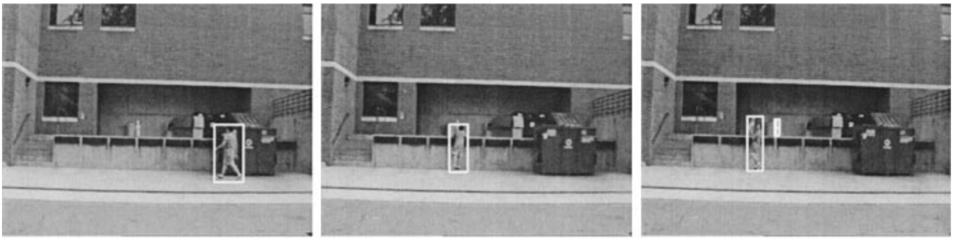

FIG. 10. A person removes an object.

histograms as probability distributions and updating them as

$$P_{t+1}(\mathbf{x} \mid i) = \beta P_t(\mathbf{x} \mid i) + (1 - \beta) P_{t+1}^{\text{new}}(\mathbf{x} \mid i), \tag{6}$$

where P_{t+1}^{new} is the probability computed using only the new image obtained at time $t + 1$, P_{t+1} is the updated probability estimate, and $0 < \beta < 1$. The sequences in this paper were processed using $\beta = 0.8$. A method for updating Gaussian mixture models is given elsewhere [23].

Color distributions were estimated in the trichromatic RGB space obtained from the frame grabber. Each channel was quantized into 16 values (4 bits). This gave a total of $16^3 = 4096$ histogram bins. This coarse quantization is easily justified if the camera produces only 4 or 5 true bits per channel.

5.1. Reasoning during Occlusions

While people are in a group there is often extensive occlusion and it is difficult to accurately segment the people from one another. However, we can still approximate their positions and obtain a partial depth ordering based on the extent to which each person is occluded.

Let A_i denote the number of pixels that were in person i's mask when that person was last observed in a group containing no other people. If nothing is known about the depth ordering of people in a group, the prior probability of a pixel corresponding to the ith person in the group can be estimated as

$$P(i) = \frac{A_i}{\sum_{j \in G} A_j}. \tag{7}$$

Each person, i, in a group has a color model $P(\mathbf{x} \mid i)$, although adaptation of these color models is suspended while a person is in a group with others because of the lack of a reliable segmentation. For each person i in a group G, and for each pixel (x, y) within the group's mask, a probability $P(\mathbf{x}_{x,y} \mid i)$ can be obtained using i's color model. Posterior probabilities can be computed by combining these probabilities with the priors:

$$P(i \mid \mathbf{x}_{x,y}) = \frac{P(\mathbf{x}_{x,y} \mid i) P(i)}{\sum_{j \in G} P(\mathbf{x}_{x,y} \mid j) P(j)}. \tag{8}$$

This can easily be implemented using histograms. Each time a group G of more than one person is formed or changes its members to form a new group, a "histogram" of posteriors

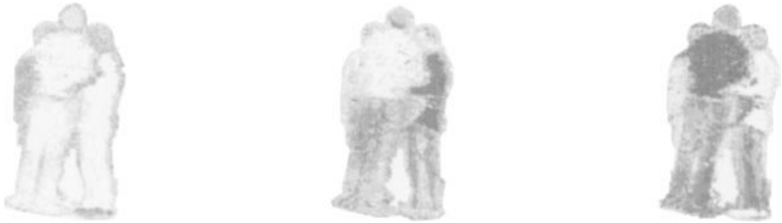

FIG. 11. Posterior probabilities for people 1, 2, and 3 respectively in frame 210. (High values are darker.)

is computed for each person in the group. This can be considered to be a generalization to multiple models of the ratio histogram idea used in [30]. The posteriors can be interpreted as follows: A high value indicates that a pixel in the group with those color values has a high probability of corresponding to an unoccluded part of that person. A low, nonzero value indicates that although the pixel could be due to the person, it is more likely to be a visible part of another person in the group. The posterior probabilities enable an estimate of the extent to which each person in the group is unoccluded to be obtained. A visibility index, $v_t(i)$, is computed for each person, i, in the group G

$$v_t(i) = \frac{\sum_{x,y} P(i \mid \mathbf{x}_{x,y})}{A_i}, \tag{9}$$

where the summation is over those pixels (x, y) in the support map for group G at time t. When $v_t(i)$ has a low value, person i is largely occluded by other people in the group. Visibility indices can be used to estimate a depth ordering of the people in the group.

Figure 11 shows the posterior probabilities computed using Eq. (8) for each person in frame 210 of the sequence in Fig. 1 (the sixth of the images shown). In this case, the shirts provide good color cues for discrimination. Persons 2 and 3 are both wearing blue jeans resulting in lower posteriors on thier legs. Figure 12 shows a plot of the visibility indices between frames 173 and 236 during which time they form a group of three. The plot correctly indicates that person 1 is heavily occluded in frame 210 while person 3 is the most visible.

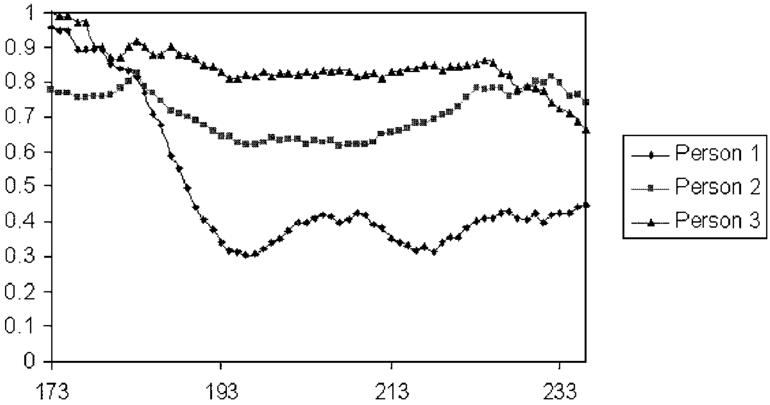

FIG. 12. Visibility indices for frames 173 to 236 of the sequence shown in Fig. 1.

5.2. *Group Bifurcation*

When a group of several people splits up to form two or more new groups, the color models of the persons in the original group are used to determine who belongs to each new group. Histogram color models are matched using the histogram intersection method proposed by Swain and Ballard for object recognition [30]. Given a histogram, H_G, computed from a newly created group G and a histogram, H_i, for person i, a normalized match value $0 \leq f(G, i) \leq 1$ is computed as

$$f(G, i) = \frac{\sum_{x,y} \min(H_G(\mathbf{x}_{x,y}), H_i(\mathbf{x}_{x,y}))}{\sum_{x,y} H_i(\mathbf{x}_{x,y})}, \tag{10}$$

where summations are over those pixels (x, y) in the support map for group G. People are allocated to groups so as to maximize the match values. The person's histogram, H_i, is used to normalize the match score. Therefore, the intersection match value is increased by a pixel in an area of the group outside the person only if (i) that pixel has the same color as a color in the model, and (ii) the number of pixels of that color in the person is less than the number of pixels of that color in the model [30].

6. DISCUSSION

Background subtraction is low level and relies entirely on local information. As such, it will never be entirely reliable but should be considered as providing useful information to intermediate level grouping processes. The scheme described here is quite robust even in unconstrained outdoor scenes. The use of adaptation is important and even allows tracking to cope with brief camera motion without complete failure. In addition, since edges are used in subtraction, this approach makes use of, and indeed favors, clutter in both the scene and the humans.

There are of course circumstances in which the tracker will fail. If two people are clothed in a very similar manner, they may be confused if they form a group and subsequently separate. Although it is possible for a person model to be erroneously initialized and tracked, this is rare because unless the regions concerned are consistently tracked over several frames, a person will not be initialized.

The tracking system described here ran successfully using several different camera and frame-grabber setups. For example, the sequences shown in this paper were captured at approximately 15 Hz using an inexpensive video camera and frame-grabber that duplicated and dropped many frames. The system also ran successfully, without the need to alter any free parameters, on sequences acquired at 60 Hz in 8-bit per channel RGB at the Keck Laboratory, University of Maryland. This helps demonstrate the robust nature of the approach. A version of this tracker that uses image processing functions optimized for MMX technology runs at approximately 8 Hz on a 500-MHz PIII PC at a resolution of 180×144 pixels.

Future work will further explore learning part-based color models for tracking people. Patterned garments might also be characterized stably using texture descriptors. Another focus of future work will be concerned with learning behavior models for person–person and person–object interactions.

REFERENCES

1. J. K. Aggarwal, Q. Cai, and B. Sabata, Nonrigid motion analysis: Articulated and elastic motion, *Comput. Vision Image Understand.* **70**, 1998, 142–156.

2. A. Baumberg and D. Hogg, An efficient method for contour tracking using active shape models, in *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pp. 194–199, 1994.

3. C. Bregler, Learning and recognizing human dynamics in video sequences, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 568–574, 1997.

4. C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8–15, 1998.

5. C. Cedras and M. Shah, Motion-based recognition: A survey, *Image Vision Comput.* **13**, 1995, 129–155.

6. Y. J. Cham and J. M. Rehg, A multiple hypothesis approach to figure tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 239–245, 1999.

7. T. Darrell, G. Gordon, M. Harville, and J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 601–609, 1998.

8. D. M. Gavrila, The visual analysis of human movement: A survey, *Comput. Vision Image Understand.* **73**, 1999, 82–98.

9. W. E. L. Grimson, C. Stauffer, R. Romano, L. Lee, P. Viola, and O. Faugeras. Forest of sensors: Using adaptive tracking to classify and monitor activities in a site, in *DARPA Image Understanding Workshop*, 1998.

10. I. Haritaoglu, D. Harwood, and L. Davis, W4: Who, when, where, what: A real-time system for detecting and tracking people, in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 222–227, 1998.

11. I. Haritaoglu, D. Harwood, and L. S. Davis, Hydra: Multiple people detection and tracking using silhouettes, in *IEEE International Workshop on Visual Surveillance*, 1999.

12. T. Horprasert, I. Haritaoglu, C. Wren, D. Harwood, L. S. Davis, and A. Pentland, Real-time 3D motion capture, in *Workshop on Perceptual User Interfaces*, 1998.

13. S. S. Intille, J. W. Davis, and A. F. Bobick, Real-time closed world tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 697–703, 1997.

14. S. Jabri, Detecting and delineating humans in video images, Master's thesis, Computer Science Department, George Mason University, Fairfax, Virginia, 1999.

15. R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*, McGraw-Hill, New York, 1995.

16. N. Johnson and D. Hogg, Learning the distribution of object trajectories for event recognition, *Image Vision Comput.* **14**, 1996, 609–615.

17. M. J. Jones and J. M. Rehg, Statistical color models with application to skin detection, Technical Report CRL 98/11, Compaq, Cambridge Research Laboratory, 1998.

18. D. Koller, J. Weber, and J. Malik, Robust multiple car tracking with occlusion reasoning, in *European Conference on Computer Vision*, pp. 189–196, 1994.

19. W. Li, H. Yue, S. Valle-Cervantes, and S. J. Qin, Recursive PCA for adaptive process monitoring, *J. Process Control*, submitted.

20. A. J. Lipton, H. Fujiyoshi, and R. S. Patil, Moving target classification and tracking from real-time video, in *DARPA Image Understanding Workshop*, 1998.

21. J. Martin, V. Devin, and J. L. Crowley, Active hand tracking, in *IEEE International Conference on Automatic Face and Gesture Recongition*, pp. 573–578, 1998.

22. S. J. McKenna and S. Gong, Recognizing moving faces, in *Face Recognition: From Theory to Applications, NATO ASI Series F* (H. Wechsler, P. J. Phillips, V. Bruce, and F. Fogelman Soulie, Eds.), Vol. 163, 1998.

23. S. J. McKenna, Y. Raja, and S. Gong, Tracking colour objects using adaptive mixture models, *Image Vision Comput.* **17**, 1999, 225–231.

24. N. Oliver, B. Rosario, and A. Pentland, A Bayesian computer vision system for modeling human interactions, in *International Conference on Vision Systems*, 1999.

25. V. I. Pavlovic, R. Sharma, and T. S. Huang, Visual interpretation of hand gestures for human–computer interaction: A review, *IEEE Trans. Pattern Anal. Mach. Intel.* **19**, 1997, 677–695.

26. S. J. Qin, W. Li, and H. Yue, Recursive PCA for adaptive process monitoring, in *World Congress of the International Federation of Automatic Control*, pp. 85–90, 1999.

27. Y. Raja, S. J. McKenna, and S. Gong, Tracking and segmenting people in varying lighting conditions using colour, in *IEEE International Conference on Face and Gesture Recognition*, pp. 228–233, 1998.

28. P. Remagnino, A. Baumberg, T. Grove, D. Hogg, T. Tan, A. Worrall, and K. Baker, An integrated traffic and pedestrian model-based vision system, in *British Machine Vision Conference*, Vol. 2, pp. 380–389, 1997.

29. C. Stauffer and W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

30. M. J. Swain and D. H. Ballard, Colour indexing, *Internat. J. Comput. Vision* **7**, 1991, 11–32.

31. C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, Pfinder: Real-time tracking of the human body, *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 1997, 780–785.