Tracking Interacting People

Stephen J. McKenna
Department of Applied Computing
University of Dundee
Dundee DD1 4HN, Scotland
stephen@computing.dundee.ac.uk

Sumer Jabri, Zoran Duric, Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030-4444, U. S. A.
{sjabri, zduric, wechsler}@cs.gmu.edu

Abstract

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 colour and gradient information is used to cope with shadows and unreliable colour cues. People are tracked through mutual occlusions as they form groups and part from one another. Strong use is made of colour information to disambiguate occlusions and to provide qualitative estimates of depth ordering and position during occlusion. Some simple interactions with objects can also be detected. The system is tested using indoor and outdoor sequences. It is robust and should provide a useful mechanism for boot-strapping and reinitialisation of tracking using more specific but less robust human models.

1. Introduction

Visual surveillance and monitoring of human activity require people to be tracked as they move through a scene. Such tracking systems can be used to learn models of activity from extended observation. These activity models can then be used to detect unusual or important events [4, 9] and to constrain tracking. Interactions with objects that are often also of interest include picking up an object, placing an object in the scene or passing an object to 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 motion of 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 would need 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. 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 dicernable. Nor is it reasonable to assume that garments are uniformly coloured. Homogeneous 'blobs' of colour will often not correspond well to body parts. However, colour 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 will vary with a person's mode of dress. The colour distributions of items of clothing are typically quite stable under rotation in depth, scaling and partial occlusion. Furthermore, colour models are easily adapted to account for gradual changes in illumination. This paper presents an efficient, colour-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 the paper is organised as follows. Section 2 briefly reviews some related work. In Section 3, a novel background subtraction technique that fuses colour and gradient information is described. A general tracking scheme which uses regions, people and groups as distinct levels of abstraction is outlined in Section 4. Section 5 describes how adaptive colour 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.

Figure 1. Frames 45, 95, 150, 180 and 195 of a sequence of people being tracked as they form groups.

2. Related Work

Several systems for tracking people have employed some form of background subtraction: Haritaoglu *et al.* used greylevel subtraction [5], Pfinder modelled pixels' colour variation using multivariate Gaussians [19], and Gaussian mixtures have also been used in a similar manner [14, 17]. 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 [15, 17]. Trackers based on 2D active shape models have been used but can only cope with moderate levels of occlusion [1, 9]. Colour 'blobs' were used in Pfinder to build a person model in a controlled indoor environment [19]. Intille *et al.*'s closed world tracking was used for example to track players on an American football pitch [7]. Haritaoglu *et al.* extended their $W4$ system to detect the heads of people in groups [5]. McKenna and Gong used a combination of motion, skin colour and face detection to track people and their faces [13]. Lipton *et al.* combined temporal differencing with template matching to track people and cars in wide-area scenes [11]. Darrell *et al.* combined depth from stereo with skin colour and face detection [3]. Bregler proposes an ambitious probabilistic framework for tracking at multiple levels of abstraction [2].

3. Adaptive Background Subtraction

We assume that the camera is stationary and that the background changes only slowly relative to the motion of 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 and three variance parameters $\sigma^2_{rcam}, \sigma^2_{gcam}, \sigma^2_{bcam}$ are estimated for the camera. Certain background pixels can violate a Gaussian assumption because of jitter or small 'micro-motions' such as leaves moving on a tree or waves on water. These changes occur on a short time-scale 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 modelled as Gaussian mixtures, for example [14, 17], it is usually not worth the added computational expense since small, isolated regions of jitter or micro-motion 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 colour 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 model is adapted on-line using simple recursive updates in order to cope with such changes. Adaptation is only performed in regions of the image which higher-level grouping processes label as background. Given a new pixel value, (r, g, b), the following updates are performed:

$$\mu_r = \alpha r + (1 - \alpha)\mu_r$$

$$\sigma^2_r = \alpha(r - \mu_r)^2 + (1 - \alpha)\sigma^2_r$$

and similarly for the g and b values. 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 $|r - \mu_r| > 3\max(\sigma_r, \sigma_{rcam})$, 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 which 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 pixels that become darker without significant chromaticity change as shadow [6, 19]. Our approach

exploits a simliar 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 chromatic change could have been caused by shadow. Chromaticity is computed as:

$$r_c = r/(r+g+b) \qquad g_c = g/(r+g+b)$$

and each pixel's chromaticity is modelled using means and variances $\mu_{rc}, \mu_{gc}, \sigma^2_{rc}, \sigma^2_{gc}$. 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 grey concrete path). In such cases, we cannot reliably tell based on zeroth-order, pixel-level, colour information whether the pixel has changed due to shadow or not. However, the use of gradient (1st order) information enables us to cope with such cases more effectively.

Gradients are estimated using the Sobel masks in x and y directions. Each pixel's gradients are modelled using gradient means (μ_{xr}, μ_{yr}) , (μ_{xg}, μ_{yg}), (μ_{xb}, μ_{yb}) and magnitude variances σ^2_{gr}, σ^2_{gg}, σ^2_{gb}. Additionally, we compute average variances $\overline{\sigma}^2_{gr}$, $\overline{\sigma}^2_{gg}$, $\overline{\sigma}^2_{gb}$ 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 $\sqrt{(r_x - \mu_{xr})^2 + (r_y - \mu_{yr})^2} > 3\max\{\sigma_{gr}, \overline{\sigma}_{gr}\}$, or if the similar test for (g_x, g_y), or (b_x, b_y) is true, then the pixel is set to the foreground. Otherwise it is set to the background.

A pixel is flagged as foreground if either chromaticity or gradient information supports that classification. A detailed description of the background subtraction method is given in [8]. 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 which would cause the greatest problems for grouping tend to have significant penumbra 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 centre image shows the copnnected components detected using an adaptive RGB background model. Much of the shadow is labelled as foreground. The rightmost image shows the result when gradient and chromaticity information are combined. Although much of the person's clothing is almost grey (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.

Figure 2. Example of background subtraction. Left: Colour image from sequence. Centre: Connected components using RGB background model. Right: Connected components using combined chromaticity and gradient background model.

3.3. Conditional hole filling

The resulting mask is likely to contain holes. Most of these holes are due to areas of low chromatic content and low texture. However, holes can also correspond to true body silhouettes when the body is in certain postures e.g. frontal view of a person with a hand on their hip. As long as such a 'hole' is not also cast in shadow, it is possible to detect it as background by examining the mask produced by the RGB subtraction described in 3.1. Therefore, conditional hole-filling is performed with reference to this mask in order to produce the final mask.

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 articulated, geometric models used for human tracking have needed to be initialised by hand in the first frame of the sequence. A robust, integrated system needs less specific models for tracking for initialisation and reinitialisation of more complex models. The system described here is not concerned with fitting models of articulated body structure or even with accurate labelling of body parts. However, it complements such approaches. We perform tracking at three levels of abstraction:

Regions Regions are connected components that are tracked consistently over time. 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 colour.

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 standard connected components algorithm is applied to the images resulting from background subtraction. Any connected components whose area is less than a threshold are discarded as noise. A temporally consistent list of tracked regions is maintained. 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 motion of regions was always small relative to their spatial extent. A region tracker is initialised for each novel region. Regions with no match are deleted. Regions can split and merge. When a region splits, all 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 3 frames, it is considered to be reliable and is subsequently considered for inclusion in people.

It is oversimplistic to assume that regions correspond to objects. People 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. A person is initialised when one or more regions that currently belong to no person together satisfy a set of rules. In order to form a person, regions must have close proximity and their projections onto the x-axis must overlap. They must together have an area larger than a threshold. Further rules based on aspect ratio, and skin colour regions can be added. Once a person is being tracked, any regions that overlap its bounding box (or alternatively its support map) are matched to that 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 in 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.

An example sequence which shows people being tracked as they form groups and split up again is shown in Figure 1. The tracking system is also able to detect some interaction with objects. If a person removes an object or deposits an object in the scene, this will give rise to a new region which splits from the person. If this region does not have significant motion and is not part of a person, then it is flagged as corresponding to an object that has just been acted upon by the person. Figures 3 and 4 show examples of detected interactions with objects.

Figure 3. A person deposits an object.

Figure 4. A person removes an object.

5. Modelling the Foreground

In order to track people consistently as they enter and leave groups, each person's appearance must be modelled. This allows people to be tracked despite the ambiguities that arise as a result of occlusion and grouping of people. A colour model is built and adapted for each person being tracked. Since people cannot be reliably segmented whilst grouped with others, person models are only adapted while a person is alone, i.e. in a group of size one.

Colour distributions have been effectively modelled for tracking using both colour histograms [12] and Gaussian mixture models [14, 16]. Both can be updated adaptively. When adaptation is not needed, a mixture can be used to generate a histogram for fast computation of probabilities. In practice these models give similar results. When the number of colour samples is small and the number of possible colour values 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 quantised colour space. For example, histograms are slightly more effective than mixture models for modelling skin colour using very large numbers of images taken from the world wide web [10]. We have used both histograms and mixture models effectively.

A histogram $H_i(\mathbf{x})$ simply counts the number of occurences 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:

$$P(\mathbf{x}|i) = \frac{H_i(\mathbf{x})}{A_i}$$

where A_i is the area of the person mask. In each frame, the model can be updated either cumulatively to model a stationary distribution or, more appropriately, adaptively to model a non-stationary distribution. Histogram models are adaptively updated by storing the histograms as probability

distributions and updating as:

$$P(\mathbf{x}|i) = \alpha P_{new}(\mathbf{x}|i) + (1-\alpha)P(\mathbf{x}|i)$$

where P_{new} is the probability computed from the new image only and $0 < \alpha < 1$. A method for updating Gaussian mixture models is given elsewhere [14].

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

5.1. Reasoning during occlusions

While people are in a larger 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 depth ordering based on the extent to which each person is occluded. For each person, i, in a group, G, and for each pixel, $\mathbf{x}$, within the group's mask, a probability $P(\mathbf{x}|i)$ is obtained using i's colour model. Posterior probabilities are then computed for each pixel in the group:

$$P(i|\mathbf{x}) = \frac{P(\mathbf{x}|i)P(i)}{\sum_{j \in G} P(\mathbf{x}|j)}$$

where,

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

This can 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 is computed for each person in the group. This can be considered to be a generalisation to multiple models of the ratio histogram idea used in [18]. The posteriors can be interpreted as follows. A high value indicates that a pixel in the group with those colour values has a high probability of corresponding to an unoccluded part of that person. A low, non-zero 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.

Whenever the composition of a group, G, changes, the posterior probabilities are summed and normalised for each person in the group:

$$\eta_i = \frac{\sum_{\mathbf{x} \in G} P(i|\mathbf{x})}{v_i}$$

where v_i is called the *visibility index* and is an estimate of the fraction of the person that is unoccluded. When a person, i, formerly in a group of size one joins the group, that person is assumed to be unoccluded so $v_i = 1.0$. At time t, the visibility index of each person in a group G of several people is estimated as:

$$v_i^t = \frac{\sum_{\mathbf{x} \in G^t} P(i|\mathbf{x})}{\eta_i}$$

When v_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 5 shows the posterior probabilities for each person in frame 150 of the sequence in Figure 1. In this case, the shirts provide good colour cues for discrimination. Persons 2 and 3 are both wearing blue jeans resulting in lower posteriors on their legs. Figure 6 shows a plot of the visibility indices between frames 110 and 174 during which time they form a group of three. The plot correctly indicates that person 1 is heavily occluded at frame 150 while person 3 is the most visible.

Figure 5. Posterior probabilities for people 1, 2 and 3 respectively in frame 150. (High values are darker)

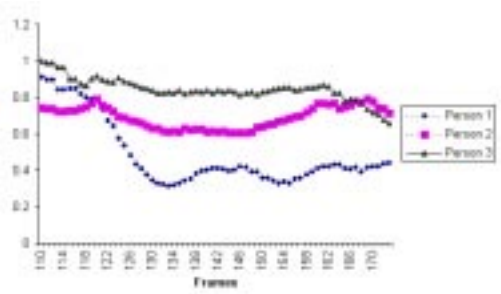

Figure 6. Visibility indices, frames 110 to 174.

5.2. Disambiguating occlusions

When a group of several people splits up to form new groups, the colour models of each person in the original group are used to determine who belongs to each new group. Histogram colour models are matched using the histogram intersection method proposed by Swain and Ballard for object recognition [18]. Given a histogram computed

from a newly created group G and a model for person i, a normalised match value $0 \leq f(G,i) \leq 1$ is computed as:

$$f(G,i) = \frac{\sum_{\mathbf{x} \in G} \min(H_G, H_i)}{\sum_{\mathbf{x} \in G} H_i}$$

People are allocated to groups so as to maximise the match values. The person's histogram is used to normalise the match score. The intersection match value is only increased by a pixel from an area of the group outside the person if (i) that pixel has the same colour as a colour in the model and (ii) the number of pixels of that colour in the person is less than the number of pixels of that colour in the model [18].

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 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.

There are of course circumstances when 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 part. Although it is possible for a person model to be erroneously initialised and tracked, this is rare because unless the regions concerned are consistently tracked over several frames, a person will not be initialised.

The tracking system described ran successfully using several different camera and frame-grabber set-ups. For example, the sequences shown in this paper were captured at approximately 15 Hz using an inexpensive video camera and frame-grabber which duplicated and dropped many frames. The system also ran successfully, without the need to alter any free parameters, on sequences acquired at 60Hz in 8-bits per channel RGB at the Keck Laboratory, University of Maryland. This helps to demonstrate the robust nature of the approach.

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

References

[1] 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*, pages 194–199, November 1994. IEEE Catalog No. 94TH0671-8.

[2] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR*, pages 568–574, 1997.

[3] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *CVPR*, pages 601–609, June 1998.

[4] 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 *Proc. of the Image Understanding Workshop*, 1998.

[5] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, when, where, what: A real-time system for detecting and tracking people. In *3rd IEEE Int. Conf. on Face and Gesture Recognition*, pages 222–227, Nara, Japan, 1998.

[6] T. Horprasert, I. Haritaoglu, C. Wren, D. Harwood, L. Davis, and A. Pentland. Real-time 3D Motion Capture. In *Proc. Workshop on Perceptual User Interfaces*, Calif., 1998.

[7] S. S. Intille, J. W. Davis, and A. F. Bobick. Real-time closed world tracking. In *CVPR*, Puerto Rico, June 1997.

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

[9] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, 1996.

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. Technical report, Compaq, Cambridge Research Laboratory, December 1998. CRL 98/11.

[11] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. In *Proc. of the Image Understanding Workshop*, 1998.

[12] J. Martin, V. Devin, and J. L. Crowley. Active hand tracking. In *IEEE Third International Conference on Automatic Face and Gesture Recongition*, Nara, Japan, April 1998.

[13] S. J. McKenna and S. Gong. Recognizing moving faces. In H. Wechsler, P. J. Phillips, V. Bruce, and F. Fogelman Soulie, editors, *Face Recognition: From Theory to Applications*, *NATO ASI Series F*, volume 163, 1998.

[14] S. J. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3-4):225–231, 1999.

[15] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. In *Proceedings of ICVS99*, Gran Canaria, Spain, January 1999.

[16] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *3rd IEEE Int. Conf. on Face and Gesture Recognition*, pages 228–233, Nara, Japan, 1998.

[17] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, pages 246–252, Fort Collins, Colorado, June 1999.

[18] M. J. Swain and D. H. Ballard. Colour indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[19] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE PAMI*, 19(7):780–785, July 1997.