Part Based Human Tracking In A Multiple Cues Fusion Framework

Qi Zhao Jinman Kang‡ Hai Tao Wei Hua‡
Department of Computer Engineering, University of California, Santa Cruz, CA, USA
{zhaoqi,tao}@soe.ucsc.edu
‡Vidient, Inc., Sunnyvale, CA, USA
{jinman,wei}@vidient.com

Abstract

This paper presents a real time video surveillance system which is capable of tracking multiple humans simultaneously. To better deal with various challenging issues such as occlusions, sharp motion changes and multi-person confusions, we propose an intelligent fusion framework where multiple cues are combined to seek the optimal objects state and more reliable cues have larger influences on the final decision. Further, part based human tracking provides a second-level information fusion in that parts with weak observability can be compensated by tracking other more visible ones, which demonstrates its effectiveness for highly articulated objects like humans.

1 Introduction

Multiple human tracking is one of the most challenging research topics in computer vision. The literature on human tracking is very large and some related methods are reviewed in the next section.

Many existing methods apply single cue, e.g. color information [4, 13, 14], for tracking. Color information is simple and effective when colors are distinguishable and no sharp illumination changes exist. However, for many real applications, color alone is not reliable due to background clutter, illumination change, low video resolution, etc. Differently, some authors fuse different visual cues [17, 21], or visual and audio data [2, 24] to more reliably solve problems in multiple human tracking. In this paper, we focus on fusing different types of visual information in a more natural way. Specifically, features we consider include motion, appearance and other local image information.

- Human motion is more complicated than motion of other objects like cars or faces. For humans, large accelerations or sudden changes in motion are common; and human articulation further aggravates the problem. Therefore human *dynamic models* are unreliable and used with caution in our method.
- Compared with motion, *appearance information* is more stable. People don't tend to change appearance from frame to frame. We build an appearance model for each body part by clustering candidate body parts, and then use these models to measure appearance similarities for body parts in later frames.
- Other useful detection or tracking modules are also incorporated into our system. For example, we build a head detector based on convolutional neural network, which helps to detect and decompose humans in each frame. When the head detection module misses possible humans, a torso detection procedure is carried out to provide reliable observation data. Such information offers strong local *image information*, which greatly compensates the irregularity of the human motion.

As pointed out by Collins *et al.* [3], the most distinctive features will change as the object moves from place to place. To address this problem, the multiple cues used in our framework perform intelligently in that their weights are adapted over time so that more reliable cues always contribute more to the final decisions.

Another novelty in this work is that in the fusion framework, human body is divided into an assembly of natural body parts to provide the second-level multiple cues to track people. By modelling and tracking each part separately, and imposing global constraint among them, the occlusion of some parts can be compensated by tracking more visible parts. Also, the problem of human articulation and fast motion are alleviated since the motion of a certain body part is obviously more regular than that of the whole human body.

2 Background and Related Work

There has been considerable work in tracking humans. A simple way to create appearance model is to model the

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

whole human body as one blob [22]. Such models are too rough for highly non-rigid objects like humans and robust tracking procedure needs to depend on other models like 3D human shape models [22]. At the other extreme are some local feature based tracking methods. In [9], the appearance is described using several typical corners instead of using the raw texture information of the whole image. Those approaches succeed in handling certain partial occlusions.

Part based human tracking methods [13, 14, 18] is a middle ground of the above two extremes. By decomposing human body into several parts, on one hand, it has its advantage on dealing with human articulation and partial occlusions; On the other hand, it is less sensitive than purely local feature based approaches. Ramanan *et al.* [13] model the 2D view of the human body using 9 colored, textured rectangles and track them based on the assumption of coherent appearance. Later they [14] developed an effective algorithm to track people by finding stylized poses. The problem is that it is computationally too expensive to apply pictorial structure model to enforce spatial relationships in each frame. Part based representations are also employed for human detection [11, 12, 19]. Those methods provide elegant results for human detection. However, completely ignoring dynamics and performing tracking by detecting parts in each frame risks high false alarms or false negatives for continuous tracking applications.

The difficulty in tracking humans can be generalized as: Firstly, occlusion in multiple human tracking, in general, remains a challenging issue. Although there are some systems proposed to deal with the occlusion problem either explicitly [16, 20, 23] or implicitly [6, 8, 10, 18], so far as I know, only small number of humans having transient occlusion can be tracked fairly reliably. Secondly, Human motion contains lots of articulated motions, and modelling the human body as a whole blob can only handle limited shape variety. Thirdly, human can move very fast and abruptly. People use the state in the current frame and a dynamic model to predict the state in the next frame. Although these predictions can be refined using image data [1, 7, 15], stable dynamics are still hard to obtain.

3 System Structure

The proposed algorithm is designed and tested in a system as Fig.1 depicts. Firstly, a background modelling algorithm is used to generate foreground mask, the result of which is sent to a head detector. Then the head detection results and camera calibration information are combined to obtain human positions and sizes as the outputs.

In this paper, we focus on the Part Based Human Tracking module, where tracking decision is an overall consideration based on part observations, dynamics, and human model constraints.

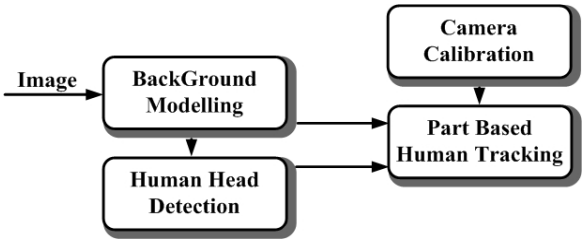

Figure 1. System Diagram.

For the following sections, in section 4, a method to obtain observation information for each human part is proposed. Section 5 introduces a tracking algorithm for each human part. In section 6, a probabilistic similarity measure is derived from the human model that combines the local features and global relationship constraints into a single equation to guide localization of parts and humans in each frame.

4 Observation

The observation information for each human part is obtained using the original image frame, the foreground mask generated by a background modelling module and the head detection results from a head detector. Any background modelling method and head detection module can be used to perform the human body decomposition, which identifies human parts, as will be introduced later. The original image provides the human part appearance information, which is modelled using color histogram calculated within the bounding box of each part.

In section 4.1, we present the human model adopted in this work, section 4.2 describes the detailed approach of human body decomposition based on the human model, and in section 4.3, part appearance models are built using the decomposition results.

4.1 Human Model

In our human model, a 2D view of the human body is modelled as a puppet of colored rectangles. 4 segments consisting of the head, the torso and two legs are used to represent the body. On one hand, to characterize the property of each part itself, orientation, width, and height for each part are modelled. On the other hand, the geometric relationships between parts are also modelled, which include relative positions between the parts and connectivity constraints, i.e., each part other than the torso needs to be *connected* to the torso, where the connectivity constraints are relaxed in that two parts are considered connected if they are close enough.

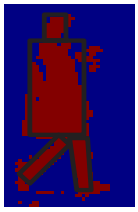

Figure 2. Human Body Decomposition.

4.2 Human Body Decomposition

Fig.2 is an illustration of the decomposition result, where the rectangles are the bounding boxes of the human parts.

4.2.1 Decomposition with head detection result

If head detection result is available, the decomposition process begins by estimating the size and position of the whole human body based on its head size and position. We obtain the bounding boxes for the torso and two legs by: Firstly, estimating the region of each part based on the human body size and position as well as relative positions of the parts provided by the human model. Secondly, on the foreground mask, extracting the major axis of each part by using PCA on the points falling in each estimated region and choosing the first principle axis. Thirdly, obtaining rectangles as the bounding boxes for each part according to its estimated region, major axis and the foreground mask.

4.2.2 Decomposition without head detection result

When head detection result is unavailable and possible humans exist, a torso detection method is proposed to compensate the observation information. Occlusion is one of the most possible reasons to cause head missing and the capability of this module to handle occlusion is shown.

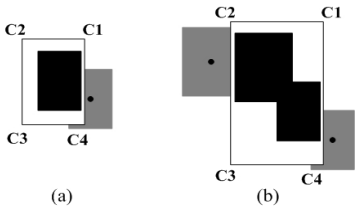

Figure 3. Illustration of Torso Detection (a) without occlusion; (b) with occlusion.

For a torso detection candidate, we search for foreground blobs (FBs) in the current frame that are overlapped with the candidate torso in the previous frame. When no occlusion happens, the case is straightforward (Fig.3(a)). The white and gray rectangles represent the bounding boxes of the FB in the current frame and the candidate torso in the previous frame, respectively. To find the best torso state in the current frame, locate the initial torso with its center at C4, the corner of the FB that is closest to the center of the torso in the previous frame (marked dot), and search for the best state using the method introduced in section 5.3. The resulting best state of the torso in the current frame is illustrated as the black rectangle. When occlusion happens, as in the case of Fig.3(b), although the FB becomes larger, the initial torso center for the right object is still C4. Therefore, the search will stop at the local maximum and not go too far to the other torso even if the two torsos are of the similar appearances.

Using the localized torso as baseline information, the head and two legs can be further localized based on the foreground mask and human model, similar as in section 4.2.1.

4.3 Part Appearance Model

The decomposition results for the first a few frames are analyzed to obtain the part appearance models. If the decomposed parts of a person satisfy human model constraints, they are kept and then each collection of one human part is sent to a mean shift procedure to run for the underlying part appearance model for that part [13].

More specifically, let a^i be the constant underlying appearance feature vector for the ith part, p_t^i, o_t^i, w_t^i and h_t^i the location, orientation, width and height of the part in frame t and Z_t^i the collection of the image patches in frame t, indexed by given position, orientation, width and height. The objective here is to achieve the maximal $P(Z_t^i|p_t^i, o_t^i, w_t^i, h_t^i, a^i)$. In essence, we are looking for a point in the domain of a^i such that for different ts, there are many $Z_t^i(p_t^i, o_t^i, w_t^i, h_t^i)$s that look like that point. By clustering the representations of the part appearances collected from the first a few frames, we are obtaining a reasonable approximation to the true max marginal of a^i.

In this work, the clustering is realized using mean shift iterations. The advantage of the mean shift clustering is that it is fast compared with many other clustering algorithms. Moreover, since the $Z_t^i(p_t^i, o_t^i, w_t^i, h_t^i)$s are obtained from the decomposed results which are checked with both the part's own estimated properties (e.g., orientation, width and height) and geometric relationships using the human model, therefore the instances of the relevant part will look like each other. This satisfies the requirement of mean shift clustering that the initialization should be good enough.

Once the appearance models are generated for all parts, they keep updated in later frames using a simple adaptive filter, which allows compensation for lighting changes, etc.

5 Human Part Tracking

The part based human tracking method uses Hidden Markov Models to model human parts, where the hidden state for each part includes location, orientation, width, height, velocity and appearance information. Each part of a person is identified by index i and its state at time t is represented by $x_t^i = (p_t^i, o_t^i, w_t^i, h_t^i, v_t^i, a_t^i)$, where p_t^i, o_t^i, w_t^i, h_t^i, v_t^i and a_t^i are the image location, orientation, width, height, velocity and appearance of the ith part, respectively. The maximum a posteriori (MAP) solution which maximizes

$$P(x_t^i|z_t^i) \propto P(z_t^i|x_t^i)P(x_t^i) \quad (1)$$

is desired, where $P(x_t^i) = P(x_0^i)$ for $t = 0$ and $P(x_t^i) = P(x_t^i|x_{t-1}^i)$ for $t > 0$.

5.1 Dynamics

The parts' predicted spatial distribution for the current image are obtained using Kalman Filter prediction.

The novelty for dynamics in this study lies in the more thorough use of the observation information: instead of using simple observations such as foreground blobs, we develop a more sophisticated approach to obtain measurements, as (will be) introduced in section 4, sections 5 (5.2 and 5.3) and section 6, to update the Kalman Filter.

5.2 Observation Likelihood

The probability $P(z_t^i|x_t^i)$ describes how the underlying state x_t^i of the ith part fits the observation z_t^i, and is defined as

$$P(z_t^i|x_t^i) = (P_A^i)_t^{(w_A^i)_t} \times (P_D^i)_t^{(w_D^i)_t} \times (P_S^i)_t^{(w_S^i)_t}, \quad (2)$$

where for each part i at time t (in the following, the indexes i and t are omitted to keep notations simple),

$P_A = \frac{1}{\alpha} e^{-\beta Distrdist(H^C, H^M)}$,

$P_D = (\frac{1}{\sigma_a\sqrt{2\pi}} e^{\frac{-(c_{cur}-c_{dec})^2}{2\sigma_a^2}})^{w_c} \times (\frac{1}{\sigma_b\sqrt{2\pi}} e^{\frac{-(\theta_{cur}-\theta_{dec})^2}{2\sigma_b^2}})^{w_\theta}$,

$P_S = \frac{1}{\sigma_c\sqrt{2\pi}} e^{\frac{-(\lambda(|diff_W/W|+|diff_H/H|+|diff_W/W-diff_H/H|))^2}{2\sigma_c^2}}$.

The observation likelihood is a weighted combination of appearance similarity, decomposition result coherence and scale similarity. w_A, w_D and w_S are weights of the three components in the likelihood computation for a certain human part, which are adaptive to different sequences and different scenarios. The underlying principle is that more reliable information and more observable parts are always assigned larger weights, e.g., appearance is more trusted when motion is irregular and torsos usually have larger weights than legs except that torsos are occluded. Besides initializations based on preliminary analysis, those weights keep updated throughout the sequence using an IIR filter.

DistrDist is a function used to compute the distance between the histogram distribution of the current image patch H^C and that of the appearance model H^M. In our implementation, the distance function is chosen as: $Distrdist(H^C, H^M) = 1 - \frac{\sum_{i=1}^m H_i^C H_i^M}{\sqrt{\sum_{i=1}^m (H_i^C)^2 \sum_{i=1}^m (H_i^M)^2}}$, where H^C and H^M are with m bins each. c_{cur} is the center of the current bounding box of the specific part, and c_{dec} is the center of the decomposed part. θ_{cur} and θ_{dec} represent the orientations of the current bounding box and the decomposed result, respectively. w_c is the weight of the location difference and w_θ the weight of the orientation difference. W and H denote the width and height of the previous bounding box. $diff_W$ and $diff_H$ are the differences in width and height of the previous and current bounding boxes, and $|diff_W/W - diff_H/H|$ penalizes the inconsistence of aspect ratio. In this algorithm, λ is chosen to be $(W + H)/2$. α, β, σ_a, σ_b and σ_c are parameters w.r.t. the variances of the information sources.

5.3 Tracking

For each human part, Kalman filter prediction and human body decomposition result serve as two threads for the initialization of the state in a new frame. Each of the two components is indispensable since motion or body decomposition module alone may not be robust under certain circumstances, e.g., when the motion changes abruptly, such as a walking person suddenly becomes static, the Kalman Filter may lose track of the person; while human body decomposition can not generate satisfactory results under occlusion.

The tracking of a single human part is essentially the process to find the best match with defined features. In this study, we propose an efficient mean-shift-alike tracking algorithm to find the best state around both initial states. In each iteration of the tracking algorithm, we search at several neighboring values and move the target window to the place with the highest posterior, instead of simply moving along the calculated moving vector through a weighted kernel computation in the mean shift iteration. This alternative has a larger recovery probability once one iteration goes incorrect. For efficiency concerns, the search for the location, orientation, width and height are realized sequentially. Starting from different initial states, the convergence points are local maxima of the posterior (Eqn.(1)) and are taken as the candidates for the specific part.

For each part, multiple candidates are kept for a human level evaluation, as will be introduced in the next section.

6 Human Body Assembling

In contrast with the model of a particular part, a model of a person should represent the generic shapes of human body. We use the posterior of the whole human configuration to describe how the underlying state of the person fits the dynamics, observations and human body model, which is a combined consideration of the posterior for each part and the model constraints for each human as a whole. The advantage of this measurement is that it couples the local and global constraints to guarantee a satisfactory match and the human body parts to be of consistent global spatial and size relationship as well as having consistent local shapes.

The human model assembling problem is to match a set of *observed* human parts, i.e., parts with high enough posteriors, against a set of human parts defined by the models. Note that this is not a one to one mapping due to occlusion, false alarm, etc. Given the frame at time t and the person is present in the image, the MAP solution for the single person tracking problem is obtained by

$$max\ P(X_t|Z_t, p), \tag{3}$$

where p denotes "person".

Applying the Bayesian rule to Eqn.(3) yields

$$max\ P(X_t|Z_t, p) \propto max\ P(Z_t|X_t, p)P(p|X_t)P(X_t),$$

where $P(p|X_t)$ is proportional to the number of identified parts, and the first and third terms can be explained as $P(Z_t|X_t, p)P(X_t) = \prod_{i=1}^{m}(P(z_t^i|x_t^i)P(x_t^i))$. Here, m is the number of identified human parts for each person, where connectivity constraints are enforced to generate those parts. $P(x_t^i)$ is obtained the same way as discussed in section 5 and $P(z_t^i|x_t^i)$ is computed using Eqn.(2).

When multiple human tracking is dealt with, the posterior becomes the product of the posteriors for single person tracking, where the derivation for each single person is identical to that for Eqn.(3). To track multiple humans simultaneously, all reasonable joint configurations for the persons are evaluated. A configuration is considered reasonable if the observed parts compose humans that fit the defined human model. The objective in the multiple human cases states as

$$max \prod_{n=1}^{N'} P((X_n)_t|(Z_n)_t, p_n)P(N'|N), \tag{4}$$

where N' is the number of persons in the current configuration and N the expected number of persons kept by the system. $P(N'|N)$ penalizes for number inconsistency.

In some sense, to obtain the MAP of a single part is to locate each individual part in the frames while seeking the MAP introduced in this section is evaluating the overall configuration of a person and performing the basic data association task. Furthermore, the data association task presented in this paper has a significant difference to many other multiple human tracking methods in that human part instead of single human is used as the basic unit for the task.

In this algorithm, we do not employ detailed human models, which may include more human parts or have more elaborate structures, e.g., the pictorial structure. Instead, we employ a coarse human model and at the same time, make full advantage of other useful but not so time-consuming modules to assist tracking.

7 Experiments

Fig.4 displays the results of tracking a person in low resolution instances. Moreover, different parts of the person share similar colors in this sequence, which imposes difficulty for many part based tracking methods using appearance information. In our fusion framework, the insufficiency, or unreliability, of certain information are compensated for by other cues. In this case, head detection results and foreground mask are more stable than appearance information, therefore are more depended by the system. Fig.4(b) shows the case of self-occlusion, where the occluded part is deleted for human body assembling.

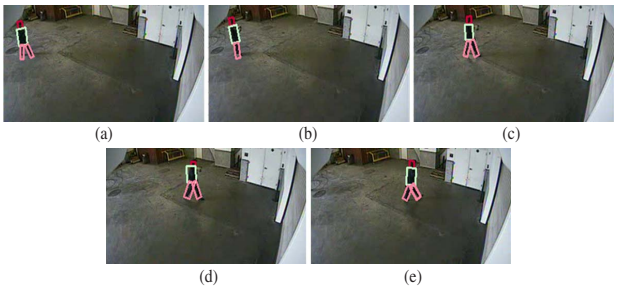

Figure 4. Tracking in low resolution.

In a second example, a simple background modelling method is applied, which tends to generate weak foreground mask information. As illustrated in Fig.5, the unsatisfactory mask throughout the sequence did not severely degenerate the tracking performance. The same effect, however, can largely perturb background subtraction based blob trackers.

Fig.6 shows another example, demonstrating the capability of the tracker to deal with inter-occlusion. When two people with similar appearance walk across each other, the decomposition module is difficult to carry out and appearance information becomes ambiguous. In such cases, the observable parts contribute more to the overall evaluation, which makes the tracking procedure constantly robust, e.g.,

Figure 5. Tracking with weak foreground mask information.

for one of the persons in Fig.6(d), only the head and torso are used to make final decision.

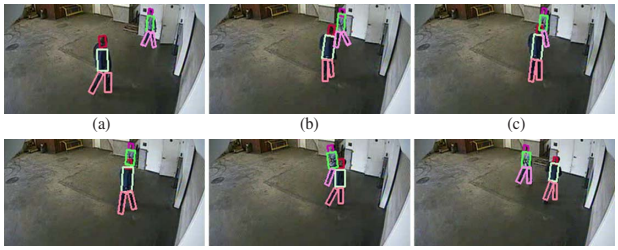

Figure 6. Tracking with occlusion. Sample frames before(a,b), in(c,d,e), and after(f) occlusion are shown.

Careful C++ implementation of the tracking algorithm allows real-time (20 fps on a 3.2 GHz PC) processing of the video streams, including the running of the head detection and background modelling modules.

8 Conclusions

This paper presents an integrated human body decomposition, part localization and human tracking vision system, where information fusion is intelligently performed. By modelling and tracking each part independently, and evaluating them using the whole human model, a natural way to robustly handle human articulation and partial occlusion is provided. It demonstrates that the system can track humans in various shapes, sizes, clothes, postures and movements. Future work will focus on the quantitative evaluation of the algorithm using public data, e.g., PETS04 dataset [5].

References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR98*, pages 8–15, 1998.

[2] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell. A probabilistic framework for multi-modal multi-person tracking. In *WOMOT03*, 2003.

[3] R. Collins, Y. Liu, and M. Leordeanu. On-line selection of discriminative tracking features. *PAMI*, 27(10):1631–1643, October 2005.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *PAMI*, 25(5):564–577, May 2003.

[5] B. Fisher. The pets04 surveillance ground-truth data sets. In *PETS04*, pages 1–5, 2004.

[6] I. Haritaoglu, D. Harwood, and L. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *VS99*, pages 6–13, 1999.

[7] D. Hogg. Model based vision: A program to see a walking person. *IVC*, 1(1):5–20, February 1983.

[8] S. Khan and M. Shah. Tracking people in presence of occlusion. In *ACCV00*, 2000.

[9] T. Mathes and J. Piater. Robust non-rigid object tracking using point distribution models. In *BMVC05*, 2005.

[10] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *CVIU*, 80(1):42–56, October 2000.

[11] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detector. In *ECCV04*, pages I: 69–82, 2004.

[12] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *PAMI*, 23(4):349–361, April 2001.

[13] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *ICCV03*, pages II: 467–474, 2003.

[14] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR05*, pages I: 271–278, 2005.

[15] K. Rohr. Incremental recognition of pedestrians from image sequences. In *CVPR93*, pages 8–13, 1993.

[16] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. In *PETS01*, 2001.

[17] N. Siebel and S. Maybank. Fusion of multiple tracking algorithms for robust people tracking. In *ECCV02*, pages IV: 373–387, 2002.

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

[19] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV05*, pages I: 90–97, 2005.

[20] Y. Wu, T. Yu, and G. Hua. Tracking appearances with occlusions. In *CVPR03*, pages I: 789–795, 2003.

[21] C. Yang, R. Duraiswami, and L. Davis. Fast multiple object tracking via a hierarchical particle filter. In *ICCV05*, pages I: 212–219, 2005.

[22] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR04*, pages II: 406–413, 2004.

[23] Y. Zhou and H. Tao. A background layer model for object tracking through occlusion. In *ICCV03*, pages 1079–1085, 2003.

[24] X. Zou and B. Bhanu. Tracking humans using multi-modal fusion. In *Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2005.