Segmentation and Tracking of Multiple Humans in Complex Situations [*]

Tao Zhao Ram Nevatia Fengjun Lv
University of Southern California
Institute for Robotics and Intelligent Systems
Los Angeles CA 90089-0273
{taozhao|nevatia|flv}@iris.usc.edu

Abstract

Segmenting and tracking multiple humans is a challenging problem in complex situations in which extended occlusion, shadow and/or reflection exists. We tackle this problem with a 3D model-based approach. Our method includes two stages, segmentation (detection) and tracking. Human hypotheses are generated by shape analysis of the foreground blobs using human shape model. The segmented human hypotheses are tracked with a Kalman filter with explicit handling of occlusion. Hypotheses are verified while they are tracked for the first second or so. The verification is done by walking recognition using an articulated human walking model. We propose a new method to recognize walking using motion template and temporal integration. Experiments show that our approach works robustly in very challenging sequences.

1 Introduction

Tracking humans in video sequences is important for a number of tasks such as video surveillance and event inference as humans are the principal actors in daily activities of interest. We consider scenarios where the camera is fixed; in this case, moving pixels can be detected fairly reliably by simply subtracting the background from the new images. In simple situations, each moving blob corresponds to a single moving object, such as a human or a vehicle; such assumption has been common in past research (e.g., [5] [8] [11], etc). However, in more complex and realistic scenarios, a single blob may contain multiple humans due to their proximity or the camera viewing angle, and also contain pixels corresponding to the shadows and reflections cast by the moving objects (See Fig.1 for examples).

Our objective in this research is to detect and track the actual walking humans in such situations in spite of the complications caused by occlusion, shadows and reflections. The tracked trajectories as well as other properties can be passed to an event recognition system [5] to perform high level interpretation of human behaviors. We have experimented with data collected on our campus with a video camera placed on the second floor (about 4m above ground) of our building capturing videos of casual passers by in front of the building under varying conditions of illumination including sunny and rainy periods.

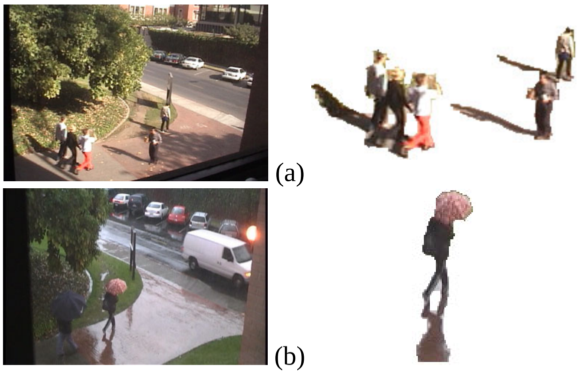


Figure 1: Example images from our dataset and two moving blobs from them.

There are many difficulties in solving this problem. We must segment a moving blob into parts that do and do not correspond to humans without knowing how many humans may be present. In tracking multiple people, the appearance of a human changes continuously due to non-rigid human motion (e.g., walking) and the changes in the viewpoints. Vehicle motion may also be present but is usually easily distinguished from human motion due to its speed and blob shape; we do not explicitly address vehicle motion in this paper.

We propose to solve the problem of human tracking under complex situations by taking advantage of various camera, scene and human models that are available and applicable for the task. We believe that the models we use are generic and applicable to a wide variety of situations. The models used are:

- *Camera model* to provide a transformation between the image and 3-D world, in conjunction with assumptions about a ground plane and upright objects, allows us to reason with physically invariant 3-D quantities,
- *Background appearance model* for motion segmentation which all the following steps are based on,

[1] This research was supported, in part, by the Advanced Research and Development Agency of the U.S. Government under contract No. MDA-908-00-C-0036.

- *Human shape model* for shape analysis (including shadow prediction) of moving blobs and for human tracking, and
- *Human articulated motion model* to recognize walking humans in the image to eliminate false hypotheses.

Most of the previous work on multi-human tracking (e.g., [5] [8] [11], etc) assumes the segmentation to individual humans is done by background subtraction and their inter-occlusion is transient. There have been some attempts at segmenting moving blobs into constituent humans in previous work such as in [7]; however, these do not consider the additional difficulties introduced by presence of shadows and reflections. [6] provides a way to handle shadows, but it uses stereo. Periodic motion analysis can be used for human detection and an overview is given in [3]. Some of the techniques are view dependent, and these techniques usually require more than one cycle of data. Furthermore, the motion of human shadow and reflection is also periodic so stronger model needs to be considered. In [10], human motion is detected by mapping the motion of some feature points to a learnt probabilistic model of joint position and velocity of different body features in a small number of frames, however, joints are required to be detected as features. Much work has been done on full body human motion tracking with an articulated human model (e.g., [1] [9], etc), but all work requires model alignment at the first frame. In motion recognition, temporal template (MEI, MHI) has been proposed in [2], however, it is view-dependent and it usually requires the actor to be in the same location.

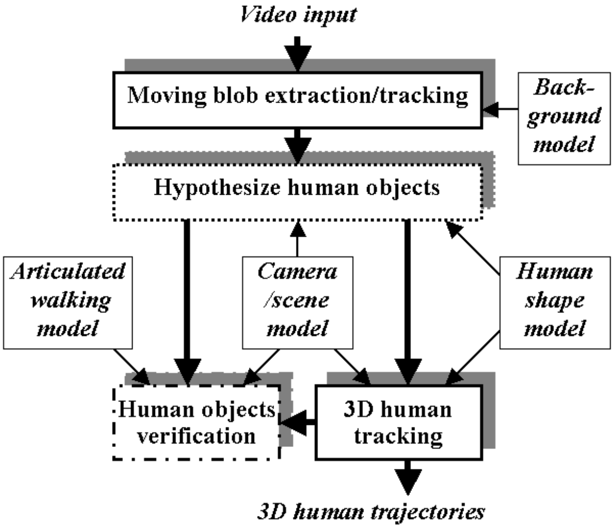

Video input
Moving blob extraction/tracking
Back-ground model
Hypothesize human objects
Articulated walking model
Camera /scene model
Human shape model
Human objects verification
3D human tracking
3D human trajectories

Figure 2: System diagram. Shaded box: program module (dotted-entry frame only, dashed-first 40 frames, solid-all frames); plain box: model; thick arrow: data flow; thin arrow: model association.

2 Our Approach Overview

The overview diagram of our approach is shown in Fig.2. First, the connected components (foreground blobs) are extracted by a background subtraction method. Our system attempts to detect humans in the extracted foreground blobs by using a "*hypothesize and verify*" approach. Human hypotheses are computed by boundary analysis (to find the head candidates) and shape analysis (to find un-explained large piece of foreground pixels). And then the hypotheses are tracked in 3D in every subsequent frame using a Kalman filter. The localization is done by template matching in a trimmed search range and the 2D positions are mapped into 3D and the trajectories are formed and filtered in 3D. However, the static information in one frame does not always yield correct hypotheses. We chose to verify the hypotheses with dynamical feature while they are being tracked. We select and verify the hypotheses by recognizing if the hypotheses exhibit a human walking pattern. In walking recognition, we use an articulated human walking model (from 3D motion captured data) to predict motion templates for a number of phases of a walking cycle, online, according to the positions and orientations of the hypotheses. Then the motion templates responses are integrated over time (for about one cycle, or 40 frames) to achieve walking recognition. The hypotheses that passed the verification are confirmed as humans and those failed are removed.

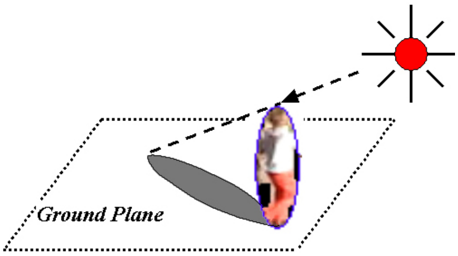


Figure 3: Human ellipsoid model and shadow prediction

3 Human Hypothesis Generation

3.1 Blob tracking

We incorporate a statistical background model [13] where each pixel in the image is an independent Gaussian color distribution. The background model is first learnt from a period where no moving object is visible and then updated by each incoming frame with the non-moving part. A single initial background frame is also sufficient to start. Our background model can be easily replaced with more complex one (e.g., multi-Gaussian [11]) if needed in extreme conditions.

The binary result after background subtraction is filtered with morphological operators. Connected components are computed, resulting in the moving blobs of that frame. (See Fig.1 for examples.) We combine the close-by blobs to avoid cases where human body is present in more than one blob; this makes the process of human segmentation more efficient.

We use a simple blob tracker to keep track of the path of and changes (split/merge) to the moving blobs. In each frame, we classify the blobs into one of *newly-created*, *disappeared*, *perfectly-matched*, *split* and *merged* by their matching with the previous frame. Our blob tracker is not perfect, for example the fast change of a blob shape can make the tracker infer it is a new blob, etc. The human tracker described later in Sec.4 is designed to work with such errors.

3.2 Scene model and calibration

The blobs extracted above are in the image domain. We use a scene model and camera calibration to map from 2D image measurements to real 3D quantities via a ground plane assumption.

First, some (about 15) 3D feature points are measured on site and their correspondences in the 2D image are given interactively. A linear calibration method [4] is used to compute the camera projection matrix. In case of known internal parameter, the estimation of external parameters can be made more robust. Camera calibration allows us to transform 2D points in the image to 3D points on a ground plane. We assume that people walk on a known ground plane. If we can locate their feet positions in the image, we can transform them to 3D scene positions. We also use the sunlight direction to compute the shadows cast on the ground by a known object, which is readily inferred from the known time, date and geographical location. (See Fig.3 for a graphical description.) However, the weather condition (e.g., sunny, cloudy) which influences whether there is shadow is given manually.

3.3 Hypothesize by shape analysis

Human hypotheses in a blob are generated by using shape information, i.e., the boundary and the vertical projection of the foreground blob. (See Fig. 4.) We attempt to explain the foreground blob with an ellipsoid (whose projection and projection of shadow are ellipses) as a coarse model of human shape.

For a camera placed several meters above the ground the head of a human has the least chance to be occluded, so we search for a point representing the top of the head in the blob boundary analysis. A point can be a head top candidate if it is the peak along the boundary within a region whose size is defined by the size of a human head. Flat peak is also allowed and the head top is defined as the center of the flat peak (Fig.4.a). We eliminate peaks that do not have sufficient number of foreground pixels below them (Fig.4.b). If a head is not overlapped with other foreground region besides its body, it will be detected most of the time and no false alarms have been found in our experiments for human blobs. After we select one head top candidate, we find its potential height by finding the first point that turns to a background pixel along a line going vertically down from the head. We do this for all the vertical lines from the head area and take the maximum, which enables finding the height of different human posture. Having the potential height, an elliptic human hypothesis is generated with the long axis (vertical) corresponding to the line from the head to the feet. If the potential height is in the valid range of human heights, we add a hypothesis of the measured height; if the potential height is over the maximum human height, we add a human hypothesis of average human height; and if the height is less than the minimum height, we add a human hypothesis of minimum human height. Once the height is fixed, the width of the ellipse is computed by the average height/width ratio of human body. The average optical flow within the elliptic mask serves as an initial estimate of the human's velocity.

After each human hypothesis is added, the un-explained foreground is updated by subtracting the human hypotheses already formed and morphological operation is performed to remove the isolated small residues (Fig.4.d). The 2D human ellipse is mapped into a 3D ellipsoid (the 2nd and 3rd axis have the same length) and its shadow cast on the ground assuming sun as the single source is also computed as shown in Fig.3. Any dark pixel in the shadow region is classified as shadow pixel and removed from the un-explained foreground. Fig.4.d shows a large proportion of the shadow pixels are removed correctly. The vertical projection of the un-explained foreground (original Fig.4.c and after update Fig.4.e) is also updated accordingly.

As mentioned above, the head won't be hypothesized if it is connected with some other part of foreground (Fig.4.d). Missing such a valid hypothesis will make some of the foreground region to be explained. The large un-explained regions will result in significant peaks in the vertical projection (Fig.4.e). From the peak in vertical projection, we search in the un-explained image for a large region and generate a hypothesis until no significant peak is found (Fig.4.f,g,h).

Since not all foreground pixels correspond to a human, there are cases where some hypotheses found do not correspond to humans. For example, when people move with their reflections, the reflections are also hypothesized as humans (as in Fig.4.i). We verify the hypotheses with dynamical features to remove such false alarms while they are being tracked, as will be described in Sec.5.

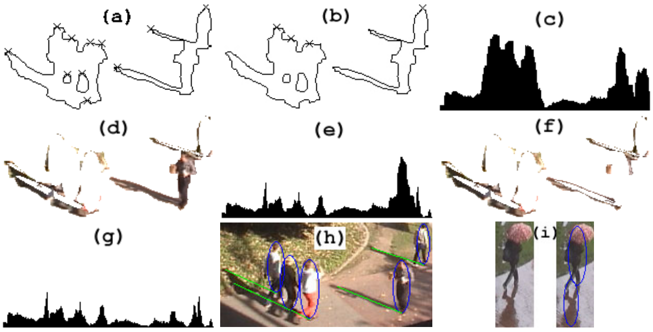


Figure 4: The process of adding human hypotheses. (Read Sec.3.3 for more detail. Note (d) (e) are un-explained foreground before morphological operation to show the removal of shadow pixels.)

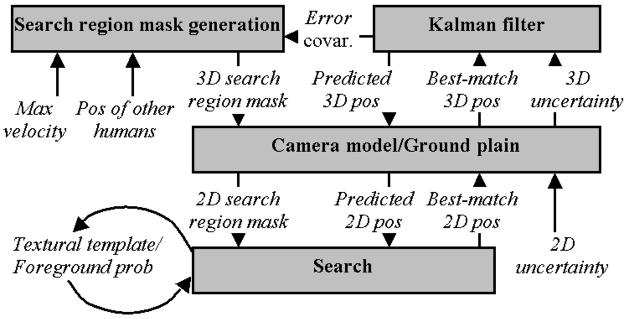


Figure 5: The diagram of 3D human tracking.

4 Human hypotheses tracking in 3D

We track each human hypothesis, whether it is verified or not since the verification is done during tracking. The human tracker is built upon the blob tracker. We call the blob that a human belongs to its owner blob. As the blob evolves over time, we also update the ownership of the humans accordingly. If a blob is perfectly matched, the ownership is inherited from last frame. If the owner blob is merged with some other blob, then the humans belonging to the previous two will share the same owner blob now. If the owner blob is split, the human(s) belongs to it will be assigned to one of the new blobs which is closer to it as its new owner blob. If a blob disappears, the human will find another blob as its owner blob according to distance.

4.1 Tracking with Kalman filter

In addition to its ownership, the accurate 3D position of each person also needs to be tracked. The diagram of the human tracker is shown in Fig.5. We track each person with a Kalman filter [12] assuming a constant velocity model (Equ.1). At each frame the position is located by template matching. The camera calibration and ground plane, again, serve as a bridge to associate 2D and 3D quantities. For each incoming frame, the 3D position of each human is predicted by the Kalman filter. The 3D position as well as the covariance matrix of the error in 3D is mapped in 2D images and a search is performed to find the best match of that person's new position. The filter is defined by:

$$\left.\begin{array}{l} X_n = AX_{n-1} + w_n \\ Z_n = BX_n + v_n \end{array}\right\} \quad (1)$$

where $X_n = [x_n, y_n, \dot{x}_n, \dot{y}_n]^T$ is the vector of the state variables, which are the 3D position and velocity. $Z_n = [x'_n, y'_n]^T$ is the measured 3D position. $U_n = [u_n, v_n]^T$ is the 2D position measured from the image directly, $Z_n = H(U_n)$ where $H(.)$ is the homography from the image to the ground plane.

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$w_n = N(0, diag(\sigma_x, \sigma_y, \sigma_{\dot{x}}, \sigma_{\dot{y}}))$ is the process noise, and $v_n = N(0, \Sigma_n^{obs})$ is the measurement noise.

The state variables are the positions and the velocities in 3D and the measurements are computed in 2D and then transformed to 3D. Assuming a fixed 2D measurement (matching) noise, the measurement noise in 3D is different when the person is at different position due to the perspective effect of the camera projection. So it is incorrect to use a fixed observation noise level throughout the tracking process. We start from the measurement uncertainty in 2D and then map it in 3D by the Jacobian of the image plane to ground plane transformation (Equ.2). This gives us good results both when humans are close to and far away from the camera.

$$\Sigma_n^{obs} = J_n * \begin{pmatrix} error_u & 0 \\ 0 & error_v \end{pmatrix} \quad (2)$$

$$J_n = \partial Z_n / \partial U_n = \partial H(U_n) / \partial U_n$$

Starting from the predicted position, a search is done to accurately locate each human. We determine the searching range with the following constraints:

- The knowledge of maximum human walking velocity,
- The a priori estimation covariance of the Kalman filter,
- And the physical occupancy constraint, i.e., two human can not overlap in 3D space.

Constraint 1 is used to specify the size of a search region mask, and all three are used to fill in the mask. Multiple humans belonging to the same blob are matched one by one starting from the one closest to the camera so that the third constraint can be enforced. Again, the mask is generated in 3D and mapped to 2D.

We use a textural template and elliptic shape mask (Fig.6, 1st, 2nd column of each human respectively) for each human to compute its match with the image. Only the pixels within the shape mask are used. Besides a textural template, in order to match in the presence of foreground/background segmentation, we also keep a map of the foreground probability P_{fg} (Fig.6, 3rd column of each human) representing the probability of each pixel of the textural template as foreground. We compute the matching error E of a pixel in the textural template with the image in the following way:

$E = P_{fg} * E_{fg} + (1 - P_{fg}) * E_{mm}$, if the corresponding pixel in the image is foreground pixel;

$E = P_{fg} * E_{mm}$, if the corresponding pixel is a background pixel;

$E = E_{occluded}$, if the pixel is occluded by other human.

where E_{fg} is the error of matching the color of the two pixels, E_{mm} is a predefined penalty for matching a foreground pixel to a background pixel and vice versa and $E_{occluded}$ is the predefined penalty for a pixel occluded by other humans.

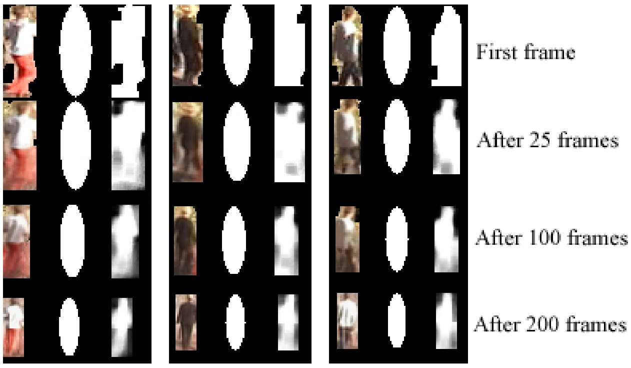


Figure 6: The evolution of textural template (1st column), shape mask (2nd column), and foreground probability (3rd column) during tracking.

The textural template is initialized at the first frame when a human is hypothesized and then updated at each subsequent frame considering human size change. The shape mask is computed automatically according to 3D height and position. In the mean while, the foreground probability map is also updated at each frame. Fig.6 shows the evolution of the templates during the tracking process. Note that the humans undergo significant size and viewpoint change.

The occlusion is explicitly handled in our tracker. There are two kinds of occlusions. Occlusion by a scene object is detected if the foreground pixel ratio within the shape mask decreases significantly, and occlusion by other human is also easily detected since we have full 3D information of all humans. We further classify the occlusion to be a partial occlusion or a full occlusion according to a visible pixel ratio threshold. In case of partial occlusion, the templates are only updated with the visible part, and in case of a full occlusion, the textural template and the Kalman filter parameters stop updating, the new position is predicted using Kalman prediction and the search range of the human is increased accordingly.

5 Human Hypothesis Verification

Human hypotheses need to be verified since non-human foreground pixels exist and it is possible that they are also hypothesized as humans (e.g., Fig.4.i). Human appearance varies significantly due to viewpoint, clothing and non-rigid motion. In some situations, even human observers have difficulty in telling the presence of humans from only a static image. Dynamical features can provide much more robust information. We observe that the motion of legs for walking people is a very salient feature, even for people of small sizes in the image. We use walking as the feature to verify the human hypotheses.

Human walking is periodic; we quantize one cycle into Q ($Q = 16$ in our implementation) phases with equal interval. We propose here a novel technique which we call motion template to represent the instantaneous motion of each phase and match with the motion the image. The motion templates are predicted by a 3D human articulated walking model, together with the camera model, the position and orientation of the hypothesis. The matches of all the frames are integrated by a straight line fitting procedure to achieve walking recognition. We set the verification process to execute for N frames ($N = 40; \sim 1.3sec$), which is the average length for a walking cycle.

5.1 Motion template

A *motion template* is the image of motion velocity (optical flow). To compute it in the image, we use a simple block-matching algorithm. In our implementation, we only compute it on foreground pixels. The motion template encodes both the shape and velocity (both amplitude and direction) of a moving object or moving parts of an object. Compared to static appearance model (e.g., edge, texture), the motion template is more distinctive since it is invariant to the texture of the object which is generally specific to individual objects such as clothing.

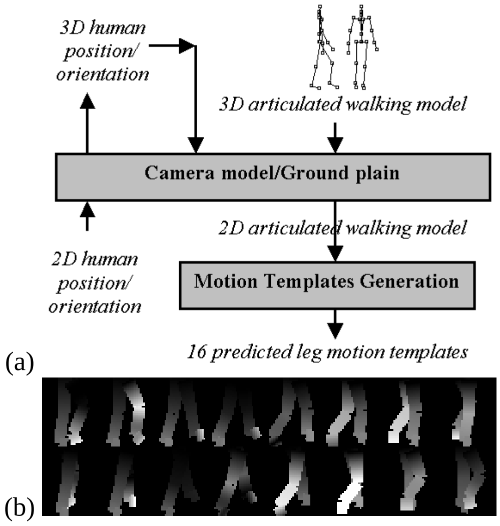
5.2 Predicting motion template from model

Motion templates are viewpoint dependent. Given the camera viewpoint and the estimated human height, position and orientation, the motion templates can be generated from the model motion data (i.e., 3D motion captured data). 3D motion captured data provides an explicit and concise representation of human articulated motion. The use of 3D motion captured data is becoming popular in human motion analysis (e.g., [9] [14], etc). The data is generally in the format of a human kinematics (articulated) model and a sequence of joint angle values over time.

We gathered a number of 3D motion captured sequences of different walking styles and computed an average walker by averaging the joint angles values after aligning them by a common starting point of a walking cycle and then temporally normalizing them into the same length. At a given phase in a walking cycle, the 3D joint positions as well as the 3D joint velocities can be computed by forward kinematics and the 2D quantities computed by camera projection. Assuming a truncated cone volumetric model for each limb segment and further simplifying the 2D projection as a quadrilateral, the motion of each pixel within the limb region is computed from the motion of the joints assuming each limb segment is a rigid body. The diagram of the prediction process is shown in Fig. 7.a and the templates generated for the upper hypothesis in Fig.9.a is shown in Fig.7.b (note only the amplitude of the flow is shown).

5.3 Walking motion recognition

A motion template only provides a description at one time instant, in order to achieve a robust verification from noisy data, we integrate multiple frames over time. We observe that the functional walking gaits of different people do not exhibit significant dynamical time warping (DTW), therefore,



3D human position/ orientation
3D articulated walking model
Camera model/Ground plain
2D articulated walking model
2D human position/ orientation
Motion Templates Generation
16 predicted leg motion templates
(a)
(b)

Figure 7: Motion template prediction. (a) the diagram; (b) motion templates predicted (only amplitude is shown; brighter means larger velocity).

the phases correspond linearly to the phases of our average walker.

$$phase_n = (K * n + phase_{init})_{MODQ} \qquad (3)$$

where n is the frame number, $phase_n$ is the phase in frame n and $phase_{init}$ is the initial phase. This linear change of the phases (Equ.3) is what we use to recognize walking.

At each frame, the correlations of the image optical flow with the Q motion templates are computed at the position we got from the human tracker. The phase that a walker is in exhibits a higher correlation value than other phases. The correlation values for all the N frames of the upper hypothesis in Fig.9 are shown in Fig.8. The peak moves in a linear fashion in a circular space ($MOD\ Q$). Note that there are two peaks in each frame made by the symmetry of left/right leg, but the correct one has larger value in most cases. The circular space complicates the application of a regular LMS straight-line fitting on the peaks if noise exists. Since we have only two parameters (K and $phase_{init}$) to estimate, a simple search technique is sufficient to find the best straight-line which maximize the sum of the correlation values in all the frames.

We perform a search on $K = tan(\theta)$ (θ is the slant angle) and $phase_{init}$, with θ changes from 45 (corresponding to maximum walking speed) to 90 degrees with a step of 1 degree and $phase_{init}$ changes from 0 to 16 with a step of 0.5. The hypothesis corresponding to a human exhibits valid walking speed (in phase space) and the summed correlation value is high. The example in Fig.9 shows the verification of two hypotheses - a human and its reflection. The K values of the valid and invalid hypotheses are 0.487 and 0.012 respectively. The line of the human is marked in blue in Fig.8, and the stick figure models corresponding to the phases are drawn according to the computed K and $phase_{init}$ in Fig.9. As can be seen, for the valid hypothesis, the computed phase matches with the image very well, while the stick-figure does not move for the invalid one.

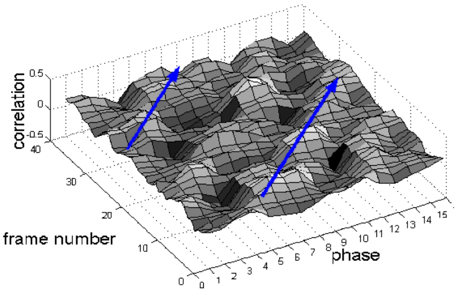


Figure 8: Motion template correlations. (Lighter color corresponds to higher correlation value.)

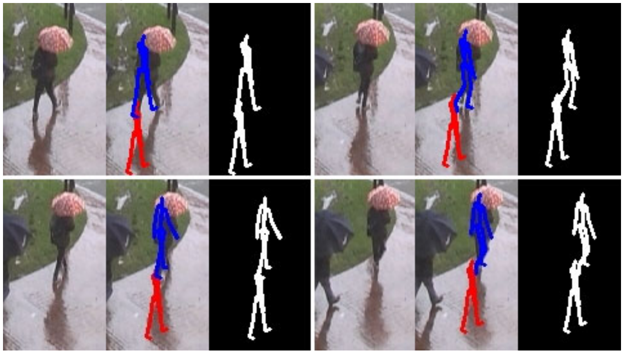


Figure 9: Key frames of a human verification example - real human (in blue) was verified while his reflection (in red) was not. For each subfigure, 1st column: original image; 2nd column: stick figure model overlaid on the image; 3rd column: stick figure model only.

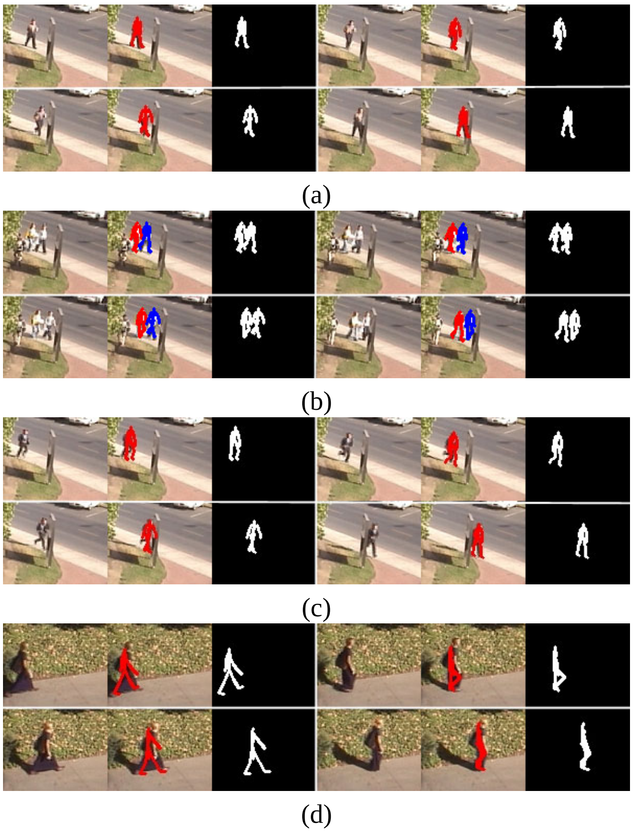
The walking pattern is a salient feature in most viewpoints except frontal/back viewpoints in which only a small amount of motion is visible in image. In these cases, since the human exhibits little motion in the image, the shape of human body can be used to verify the hypotheses. If a human's legs are highly occluded, the verification cannot be performed.

6 Results and Evaluation

We performed experiment on the proposed methods on a set of real-life data which we collected on campus. Some of the results are shown below [2].

6.1 Human tracking results

We have tested the human tracker on a number of sequences and got satisfactory results. Due to the space limit, we only show the tracking result of one representative sequence in Fig.11. The sequence includes human walking in group of 3 very closely, human passing-by each other and single/multiple human passing an obstacle. All humans were tracked successfully and the positions were estimated fairly


(a)

(b)

(c)

(d)

Figure 10: More hypothesis verification results. (1st column: original, 2nd column: stick-figure overlaid, 3rd column: stick-figure only)

accurately. As the key frames (as well as the MPEG movie) show, the bounding boxes of foreground blobs split and merge frequently while humans move smoothly. The search region size changes a lot when humans are in different positions and grows when passing the occluding object.

6.2 Hypothesis verification results

Besides the example shown in Fig.9, a wide variety of verification results were collected as shown in Fig.10 in the same format. In our experiments, we set the verification process to start 4 frames after the hypothesis is formed to get a better estimate of its orientation. Even if so, the estimate is still about 10 degrees off the real value, but it is hardly seen from the result.

Fig.10.(a) shows an example where a human enters from the far end of the scene. (b) shows an example of two close-by walkers. (c) shows a human running into the scene. In all the three examples, the human height is about 25 pixels in the image and there are frames in which the walkers are occluded by the map board. (d) shows an example of lady in long dress. In all the sequences including the running one, the phase alignment is very accurate. The results are very promising since the small image size is difficult even for human observers and our method was not even confused by left/right leg ambiguity.

6.3 Evaluation

We have tested the system a number of video segments with human activity totaling 500 seconds in time. 45 distinct people appeared in the data. Over 95% of the humans in each frame are correctly detected and tracked. In the data, there are 11 cases where people walk side by side; 14 cases where people pass by each other; and 23 cases where people are temporarily occluded (partially or completely) by scene objects. The false alarms and missed detections before and after verification are shown below:

	Before verification	After verification
False alarms	12	2
Missed detections	1	6

As can be seen, walking verification reduced the number of false alarms from 12 down to 2, however, it also rejects an extra 5 real humans. This happens mostly when the walkers are walking towards or away from the view direction so that the motion of the leg is not as salient as when walking in other directions.

7 Conclusion and future work

We have described our work on segmentation and tracking of multiple human in complex situations. Our approach can successfully handle shadow, reflection, multiple human in one moving blob, and occlusion.

The contribution of our work lies in the employment of appropriate models and knowledge to robustly solve a difficult and useful problem. We use a background appearance model to focus our interest by throwing away the static regions. We take advantage of the camera and scene (ground plane) model to get 3D quantities from 2D measurements and use the invariant 3D quantities back in 2D analysis; it, as can be seen in the diagrams, has been serving as a central bridge in our processing. The elliptic (ellipsoid in 3D) human shape model gives reasonable approximation of human shape with a low dimensional parametric form. The articulated walking model provides a compact representation of human walking pattern. Nevertheless its simplicity, we have seen its generality on various walkers even a runner.

We proposed the use of motion (optical flow) template as an appearance model in the presence of motion. We also proposed a simple technique to recognize linear (circular) motion. The motion template, combined with temporal integration gives very robust recognition results. It may be used for more complex motion recognition tasks with other integration techniques.

The most appealing point of our system is that it does not require intermediate steps, such as the foreground extraction,

[2] The MPEG movies of the tracking and verification results are available at http://iris.usc.edu/˜taozhao/arda/tracking.html.

blob tracker, optical flow computation the estimation of orientation of humans to be perfect. We believe this is an important point in a real-life application.

The work can be improved and extended in the following aspects. A better optical flow algorithm needs to be devised. More study can be done to determine the minimum number of frames needed for verification. We need to automatically decide the time to do human detection when a human or a group of human entirely enters the image. Different techniques might be used to verify humans walking along the camera view angle. The measurements (e.g., human height) can be refined during tracking, instead of fixed by the value of detection.

References

[1] C. Bregler and J. Malik, Tracking people with Twists and Exponential Maps, *CVPR'98*.

[2] A. F. Bobick, J. W. Davis, The Recognition of Human Movement Using Temporal Templates, *IEEE Trans. on PAMI*, vol. 23, no. 3, 2001.

[3] R. Cutler and L. S. Davis, Robust Real-Time Periodic Motion Detection, Analysis, and Applications, *IEEE Trans. on PAMI*, vol. 22, no. 8, 2000.

[4] D. Forsyth and J. Ponce, *Computer Vision : A Mordern Approach*, Prentice-Hall, 2001.

[5] S. Hongeng and R. Nevatia, Multi-Agent Event Recognition, *ICCV'01*.

[6] I. Haritaoglu, D. Harwood and L. S. Davis, W4S: A Real-Time System for Detecting and Tracking People in 2 1/2 D, *ECCV'98*.

[7] S. Haritaoglu, D. Harwood and L. S. Davis, W4: Real-Time Surveillance of People and Their Activities, *IEEE Trans. on PAMI*, Vol. 22, No. 8, 2000.

[8] R. Rosales and S. Sclaroff, 3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions, *CVPR'99*.

[9] Sidenbladh, H., Black, M.J. and Fleet, D.J., Stochastic tracking of 3D human figures using 2D image motion, *ECCV'00*.

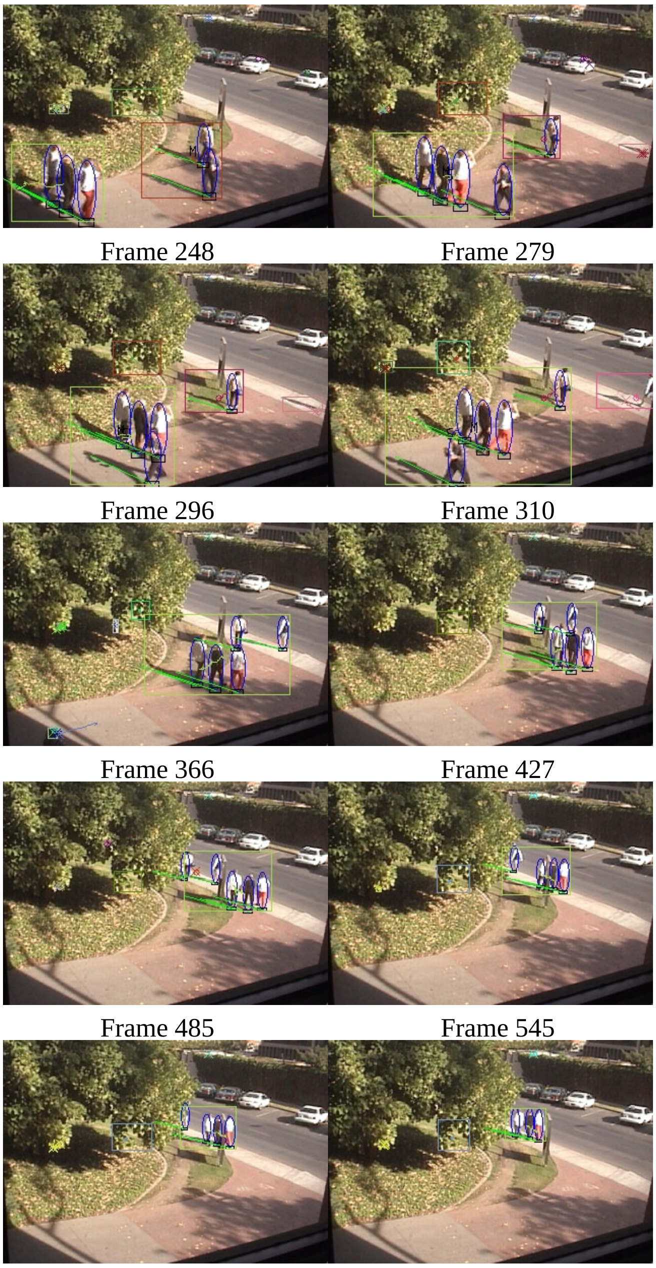
[10] Y. Song, X. Feng, P. Perona, Towards Detection of Human Motion, *CVPR'00*.

[11] C. Suauffer, W. E. L. Grimson, Learning Patterns of Activity Using Real-Time Tracking, *IEEE Trans. on PAMI*, vol. 22, no. 8, 2000.

[12] G. Welch, G. Bishop, An Introduction to the Kalman Filter, *TR-95-041, Dept. of Computer Science, Univ. of North Carolina at Chapel Hill*.

[13] Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P. Pfinder: real-time tracking of the human body *IEEE Trans. on PAMI*, Vol. 19, No. 7, 1997.

[14] T., T. Wang, H. Shum, Learning highly structured motion model for 3D figure tracking, *ACCV'02*.



Frame 248 Frame 279

Frame 296 Frame 310

Frame 366 Frame 427

Frame 485 Frame 545

Frame 563 Frame 593

Figure 11: Key frames of tracking result of one video sequence. For blob tracker, the color of bounding boxes reflects the temporal correspondence, trails are also shown in same color, "M" means merge and "S" means split. For human tracker, human objects are shown in blue ellipses, shadow direction is shown in green lines (due to the incapability of drawing an ellipse in arbitrary direction), search mask size is shown at the feet of each human.