A System for Dissecting the Video for Tracing Multiple Humans in Multifaceted Situation

M.Hemalatha, Ph.D
Head, Dept of SoftwareSystems, Karpagam University Coimbatore

V.Vinodhini,
Assistant Professor
Dr.N.G.P Arts And Science College,Coimbatore-48

B.Sivaranjani,
Assistant Professor
Dr.N.G.P Arts And Science College,Coimbatore-48

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. This 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.

General Terms
Human Tracking, Segmentation, Kalman Filter, Motion Template.

1. INTRODUCTION
Tracking humans in video sequences is important for 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 . 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 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 to perform high level interpretation of human behaviors.

There are many difficulties in solving this problem. This 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. 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. 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.

Human shape model for shape analysis (including shadow prediction) of moving blobs and for human tracking.

Human articulated motion model to recognize walking humans in the image to eliminate false hypotheses. Most of the previous work on multi-human tracking assumes the segmentation to individual humans is done by background subtraction and their inter-occlusion is transient. Periodic motion analysis can be used for human detection. 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. Much work has been done on full body human motion tracking with an articulated human model but all work requires model alignment at the first frame.

2. APPROACH OVERVIEW
First, the connected components (foreground blobs) are extracted by a background subtraction method. The 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.

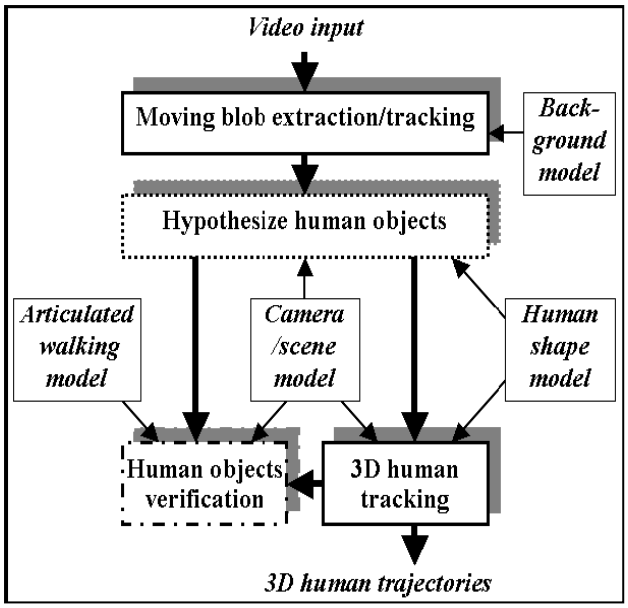

Fig 1: The system diagram. Shaded box: program module; plain box: model; thick arrow: data flow; thin line: model association.

Chose to verify the hypotheses with dynamical feature while they are being tracked. To select and verify the hypotheses by recognizing if the hypotheses exhibit a human walking pattern. In walking recognition, 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.

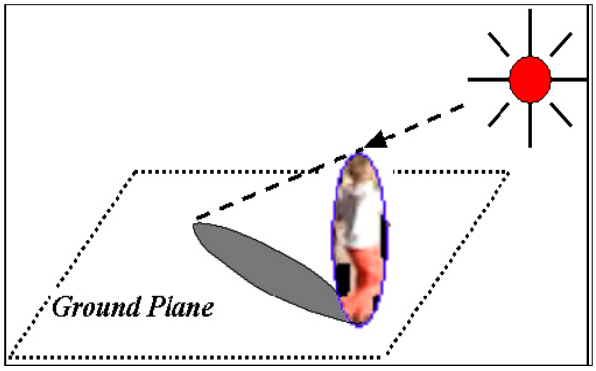

Fig 2: Human ellipsoid model and shadow prediction

2.1 Human Hypothesis Generation

Blob tracking

Iincorporate a statistical background model 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. The background model can be easily replaced with more complex one (e.g., multi-Gaussian 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. To 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. To use a simple blob tracker to keep track of the path of and changes (split/merge) to the moving blobs. In each frame, to classify the blobs into one of newly-created, disappeared, perfectly-matched, split and merged by their matching with the previous frame. The 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.

2.2 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. 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. To observe that the motion of legs for walking people is a very salient feature, even for people of small sizes in the image. To use walking as the feature to verify the human hypotheses.

2.3 Motion template

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. To set the verification process to execute for frames which is the average length for a walking cycle.

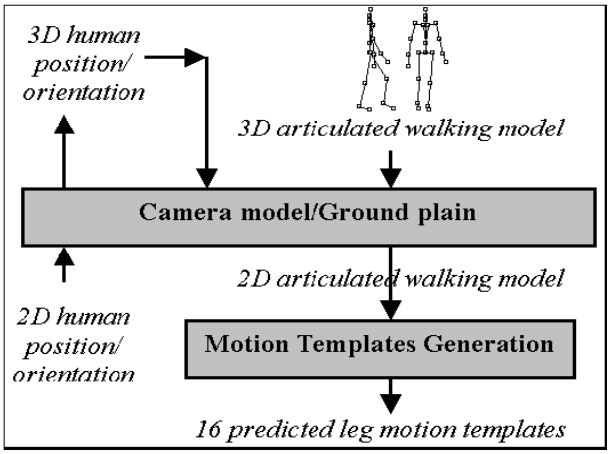

Fig 3: Motion template

A *motion template* is the image of motion velocity (optical flow). To compute it in the image, to 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.

3. SEGMENTATION AND T RACKING OF M ULTIPLE H UMANS

3.1 Background Model, Camera/Scene Model, and Human Shape Model

To incorporate a statistical background model where the color of each pixel in the image is modeled by a Gaussian distribution. The background model is first learnt in a period where there are no moving objects in the scene and then updated for each incoming frame with the on moving pixels. A single initial background frame insufficient to start. The background model can be easily replaced with a more complex one (e.g., one with a multi- Guassian model or one which can start with moving objects in the scene if needed. Change detection is performed on each incoming frame. The pixels whose values are sufficiently different from the corresponding background models are classified as fore-ground pixels. The binary map is filtered with a median filter and the morphology close operator to remove isolated noise, resulting in the foreground mask F. Connected components are then computed, resulting in the moving blobs (or, simply, blobs) of that frame. This allows a larger coverage and less occlusion, especially avoiding the situation where the entire scene is occluded by one object. Such a setup is also in accordance with most commercial surveillance systems. To compute the camera calibration, the traditional approach requires enough 3Dfeaturepoints(? 6pointswith ? 2 of them out of a plane) and their corresponding image points.

A linear calibration method described works satisfactorily if the selected points are distributed evenly in the image. If the number of feature points is not enough or measurement of 3D points is not possible, methods based on the projective invariance (e.g., vanishing points) can be used , walking in more than one direction can provide enough information for an approximate camera calibration. Both methods have been used in our experiments. To assume that people move on a known ground plane. The camera model and the ground plane together serve as a bridge to transform 2D and 3Dquantities.Three-dimensional quantities can be projected into 2D quantities by the camera model. The camera model and the ground plane define a transformation (i.e., a homography) between the points on the image plane and the points on the ground plane. The measurements of the objects (such as position, velocity, and height) in the image can be transformed into 3D. Sometimes, only to know the position of a human's head instead of his/ her feet. Then, the transformation can be carried out approximately by assuming that the humans are of an average height. The transformation degenerates when the projection of the reference plane is (or close to) a line in the image, i.e., when the optical axis is on the reference plane. Such a case does not occur in our camera setup. To model gross human shape by a vertical 3D ellipsoid. The two short axes are of the same length and have a fixed ratio to the length of the long axis. The parameters of an object include its position on the ground plane and its height. Assuming an ellipsoid is represented by a 4 by 4 matrix, Q, in homogenous coordinates, its image under camera projection P (a 3 by 4 matrix) is an ellipse, represented by a 3 by 3 matrix, C. An object mask M is defined by the pixels inside the ellipse. The 3D human shape model also enables geometric shadow analysis.

3.2 Segmenting Multiple Humans

Attempt to interpret the foreground blobs with the ellipsoid shape model. Human hypotheses are generated by analyzing the boundary and the shape of the foreground blobs.

4. RESULTS AND EVALUATION

4.1 Human tracking results

To test the human tracker on a number of se-quences and got satisfactory results. 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 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.

4.2 Evaluation

To test 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 ob- jects. The false alarms and missed detections before and after verification are shown below: Before verification After verification

Table 1: Before verification After verification

	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

Besides observing the results visually, we evaluate the performance of the system quantitatively on a dataset. The sequences in the dataset are selected to be heterogeneous. They are obtained from different sources , captured on nine distinct sites, with the camera tilt angle ranging from 5 to 40 degrees and image height of a typical human ranging from about 25 pixels to about 80 pixels. The total length of the data is 61,890 frames (35 minutes) and the total number of humans

that appear is 520 (counted by a human observer). The number of human-frames (i.e., the summation of the frames of each human) is 243,804 (135 minutes).

The performance of human segmentation and tracking is affected by the complexity of the data. Generally, the more the occlusion, the more challenges the data pose to the algorithms. It is known that track drifts or switches are more likely to happen when a few people walk in a group or pass-by each other. To describe the complexity of the dataset by the number of the events challenging to the system: "passing-by each other," "passing an obstacle," and "walking in a

group." A "passing-by" means that two or more humans cross each other with the human farther away from the camera partially or completely occluded. A "passing an obstacle" means that a human moves behind a structure (e.g., a tree or a pole) in the scene. A "walking in a group" means that two or more humans walk together closely (e.g., side by side) and persistently.

Quantitative evaluation of multi-object tracking is more difficult than object detection, object recognition or single object tracking due to the complex behaviors created by the system. These behaviors need various statistics to be computed in order to be meaningful to different applications.

5. Conclusion and future work

Here the work is described 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. To use a background appearance model to focus our interest by throwing away the static regions. The 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, blob tracker, optical flow computation the estimation of orientation of humans to be perfect.

This is an important point in a real-life application because 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. 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.

6. 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] W. Hu, T. Tan, L.Wang, and S. May bank, "A Survey on Visual Surveillance of Object Motion and Behaviors", IEEE Transactions on Systems, Mann, and Cybernetics, Part C: Applications and Reviews, vol. 34, No. 3, Aug. 2004, pp. 334-352.

[10] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, 2000, pp.142–149.

[11] C. Stauffer and W.E.L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug. 2000.

[12] H. Tao, H.S. Sawhney, and R. Kumar, "A Sampling Algorithm for Tracking Multiple Objects," Proc. IEEE Workshop Vision Algorithms, 1999.

[13] H. Tao, H.S. Sawhney, and R. Kumar, "Object Tracking with Bayesian Estimation of Dynamic Layer Representations," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, Jan. 2002.

[14] A.M. Tekalp, Digital Video Processing. Prentice Hall, 1995. [44] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, July 1997.

[15] T. Zhao, R. Nevatia, and F. Lv, "Segmentation and Tracking of Multiple Humans in Complex Situations," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 194-201, 2001.

[16] T. Zhao and R. Nevatia, "3D Tracking of Human Locomotion: A Tracking as Recognition Approach," Proc. Int'l Conf. Pattern Recognition, vol. 1, pp. 546-551, 2002.

[17] T. Zhao, "Model-Based Segmentation and Tracking of Multiple Humans in Complex Situations," PhD thesis, Univ. of Southern California, Los Angeles, 2003.

[18] A. Prati, R. Cucchiara, I. Mikic, and M.M. Trivedi, "Analysis and Detection of Shadows in Video Streams: A Comparative Evalua- tion," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 571-576, 2001.

[19] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Slected Applications in Speech Recognition," Proc. IEEE, vol. 77, no. 2, 1989.

[20] K. Rohr, "Towards Model-Based Recognition of Human Move- ments in Image Sequences," CVGIP: Image Understanding, vol. 59, no. 1, pp. 94-115, 1994.

[21] R. Rosales and S. Sclaroff, "3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 117-123, 1999.

[22] H. Sidenbladh, M.J. Black, and D.J. Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion," Proc. European Conf. Computer Vision, pp. 702-718, 2000.

[23] N.T. Siebel and S. Maybank, "Fusion of Multiple Tracking Algorithm for Robust People Tracking," Proc. European Conf. Computer Vision, pp. 373-387, 2002.

[24] Y. Song, X. Feng, and P. Perona, "Towards Detection of Human Motion," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 810-817, 2000.

[25] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," J. Basic Eng., vol. 82, pp. 35-45, 1960.

[26] P. Kornprobst and G. Medioni, "Tracking Segmented Objects Using Tensor Voting," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 118-125, 2000.

[27] N. Krahnstover, M. Yeasin, and R. Sharma, "Towards a Unified Framework for Tracking and Analysis of Human Motion," Proc. IEEE Workshop Detection and Recognition of Events in Video, 2001.

[28] D. Liebowitz, A. Criminisi, and A. Zisserman, "Creating Archi- tectural Models from Images," Proc. EUROGRAPH Conf., vol. 18, pp. 39-50, 1999.

[29] A.J. Lipton, H. Fujiyoshi, and R.S. Patil, "Moving Target Classification and Tracking from Real-Time Video," Proc. DARPA IU Workshop, pp. 129-136, 1998.

[30] F. Lv, T. Zhao, and R. Nevatia, "Self-Calibration of a Camera from a Walking Human," Proc. Int'l Conf. Pattern Recognition, vol. 1, pp. 562-567, 2002.

[31] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking Groups of People," Computer Vision and Image Under- standing, vol. 80, no. 1, pp. 42-56, 2000.