A STUDY OF MULTIPLE HUMAN TRACKING FOR VISUAL SURVEILLANCE

Shalini Agarwal, Shaili Mishra
Department of CS, Banasthali University, Rajasthan

ABSTRACT

Visual surveillance has become very active research topic in computer vision. This paper deals with the problem of detecting and tracking multiple moving people in a static background. Detection of foreground object is done by background subtraction. Tracking multiple humans in complex situations is challenging. The difficulties are tackled with appropriate knowledge in the form of various models in our approach. Human motion is decomposed into its global motion and limb motion. Our objective in this paper is to segment multiple human objects and track their global motion in complex situations where they may move in small groups, have interocclusions, cast shadows on the ground, and reflections may exist.

KEYWORDS: *Background subtraction Method, Blobs, Optical Flow, Multiple-human segmentation, multiple-human tracking, human locomotion model.*

I. INTRODUCTION

Automatic visual surveillance in dynamic scenes has recently got a considerable interest to researchers. Technology has reached a stage where mounting video camera is cheap causing a widespread deployment of cameras in public and private areas. It is very costly for an organization to get their surveillance job done by humans. Beside cost, other factors such as accuracy, negligence makes manual surveillance inappropriate. So, automatic visual surveillance have becomes inevitable in the current scenario. It will allow us to detect unusual events in the scene and warrant the attention of security officers to take preventive actions. The purpose of visual surveillance is not to replace human skill and intuition power but is to assist human for smooth running of the security system.

The object can be represented as:-

- **Points***:* The object is represented by a point, that is, the centroid (Figure 1(a)) In general, the point representation is suitable for tracking objects that occupy small regions in an image.
- **Primitive geometric shapes:** Object shape is represented by a rectangle, ellipse (Figure 1 (c), (d)), etc. Though primitive geometric shapes are more suitable for representing simple rigid objects, they are also used for tracking no rigid objects.
- **Object silhouette and contour**: Contour representation defines the boundary of an object (Figure 1(g), (h)). The region inside the contour is called the silhouette of the object (see Figure 1(i)). Silhouette and contour representations are suitable for tracking complex no rigid shapes.
- **Articulated shape models:** Articulated objects are composed of body parts that are held together with joints. For example, the human body is an articulated object with torso, legs,

hands, head, and feet connected by joints. In order to represent an articulated object, one can model the constituent parts using cylinders or ellipses as shown in Figure 1(e).

- **Skeletal models:** Object skeleton can be extracted by applying medial axis transform to the object silhouette. This model is commonly used as a shape representation for recognizing objects. Skeleton representation can be used to model both articulated and rigid objects (Fig 1(f)).[1][2]

It is difficult to get a background model from video because background information keeps always changing by factors such as illumination, shadow.[3] So a static background is assumed. Well-known background subtraction method is used for detecting moving object, because it gives maximum number of moving pixels in a frame.

Object tracking methods can be divided into 4 groups, they are:

- Region-based tracking
- Active-contour-based tracking
- Feature-based tracking
- Model-based tracking

It is not so easy because of some of the problems, which generally occur during tracking. Occlusion handling problem i.e. overlapping of moving blobs has to be dealt carefully[6][7]. Other problems like lighting condition, shaking camera, and shadow detection, similarity of people in shape, color and size also pose a great challenge to efficient tracking

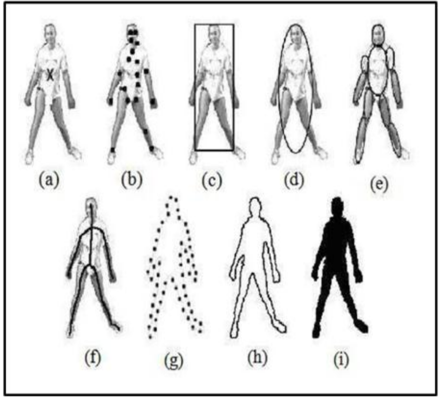


Fig 1: Object Representation

The rest of the paper is organized as follows: section II gives a survey of techniques used for human tracking in surveillance system .Section III theoretical Background about tracking system. Section IV presents some of the problems occurs in existing technologies and problem formulation. Section V presents solution approach .In Section VI we remove problem of occlusion in multiple human tracking system. Conclusion and future work is given in section VII.

II. RELATED WORK

Most of the work on tracking for visual surveillance is based on change detection [44][36][40][15][13][11][21][38] or frame differencing [23] if the camera is stationary. Additional stabilization is required if the camera is mobile [7][42]. These methods usually infer global motion only and can be roughly grouped as follows:

- Perceptual grouping techniques are used to group the blobs in the spatio-temporal domain as in Cohen and Medioni [7] and Kornprobst and Medioni [20]. However, these methods still suffer from the deficiencies of blob-based analysis discussed earlier. In Lipton et al. [23], a moving blob is classified into a single human, multiple-human or a vehicle accord-ing to its shape. However, the positions of the people in a multihuman blob is not inferred.
- Some work (Rosales and Sclaroff [36], Elgammal and Davis [11], and McKenna et al. [25], etc.) assumes people are isolated when they enter the scene so that an appearance model can be initialized to help in tracking when occlusion happens. These methods cannot be applied where a few people are observed walking together in a group.
- Some methods try to segment multiple people in a blob. The W^4 system [15] uses blob vertical projection to help segment multiple humans in one blob. It only applies to data where

multiple people distribute horizontally in the scene ("step on one's head" does not happen, usually from a ground level camera). It handles shadows by use of stereo cameras [14]. Siebel and Maybank [38] extend the Leeds human tracker [1] by the use of a head detection method similar to the approach taken in our system.

- Tao et al. [41] and Isard and MacCormick [18] track multiple people using the CONDENSATION algorithm [17]. The system in [18] also uses a human shape model and the constraints given by camera calibration. It does not involve any object-specific representation; therefore, the identities of humans are likely to be confused when they overlap. Besides, the performance of particle filter is limited by the dimensionality of the state space, which is proportional to the number of objects.

Other related work includes Tao et al. [42] which use a dynamic layer representation to track objects. It combines compact object shape, motion, and appearance in a Bayesian framework. However it does not explicitly handle occlusion of multiple objects since it was designed mainly for airborne video.

Much work has been done on estimating human body postures in the context of video motion capture (a recent review is available in [26]). This problem is difficult, especially from a single view because 3D pose may be under constrained from one viewpoint. Most successful systems (e.g., [9]) employ multiple viewpoints, good image resolution, and heavy computation, which is not always feasible for applications such as video surveillance. Use of constrained motion models can reduce the search space, but it only works on the type of motion defined in the model. Rohr [35] describes pioneering work on motion recognition using motion captured data. In each frame, the joint angle values are searched for on the motion curves of a walking cycle. Results are shown only on an isolated human walking parallel to the image plane. Motion subspace is used in Sidenbladh et al. [37] to track human walking using a particle filter. Both [35] and [37] operate in an online mode. Bregler [4] uses HMMs (hidden Markov models) to recognize human motion (e.g., running), but the recognition is separated from tracking. Brand [3] maps 2D shadows into 3D body postures by inference in an HMM learnt from 3D motion captured data, but the observation model is for isolated objects only. In Krahnstover et al. [21], human tracking is treated as an inference problem in an HMM; however, this approach is appearance-based and works well only for the viewpoints for which the system was trained. For motion-based human detection, motion peri-odicity is an important feature since human locomotion is periodic; an overview of these approaches is given in [8]. Some of the techniques are view dependent, and usually require multiple cycles of observation. It should be noted that the motion of human shadow and reflection is also periodic. In Song et al. [39], human motion is detected by mapping the motion of some feature points to a learned probabilistic model of joint position and velocity of different body features, however, joints are required to be detected as features. Recently, an approach similar to ours has been proposed by Efros et al. [10] to recognize actions. It is also based on flow-based motion description and temporal integration.

III. THEORETICAL BACKGROUND

3. 1 Object Segmentation

Most of the work on foreground objects segmentation is based on three basic methods, namely frame differencing, background subtraction and optical flow. Only background subtraction requires modeling of background. It is faster than other methods and can extract maximum features pixels. It uses a hybrid of frame differencing and background subtraction for effective foreground segmentation. A considerable amount of work has been done on modeling dynamic background. Researchers usually use Gaussian, a mixture of Gaussian, kernel density function or temporal median filtering techniques for modeling background. Assuming that surveillance is taken at the scenario, which is static background. Object extraction i.e. foreground segmentation is done by Background Subtraction. Building a representation of the scene, called the background model and then finding deviations from the model for each incoming frame can achieve object detection. Any significant change in an image region from the background model signifies a moving object. Usually, a connected component algorithm is applied to obtain connected regions corresponding to the objects. This process is referred to as the background subtraction [30].

3.2 Background Subtraction

Background subtraction is a computational vision process of extracting foreground objects in a particular scene. A foreground object can be described as an object of attention which helps in reducing the amount of data to be processed as well as provide important information to the task under consideration. Often, the foreground object can be thought of as a coherently moving object in a scene. We must emphasize the word coherent here because if a person is walking in front of moving leaves, the person forms the foreground object while leaves though having motion associated with them are considered background due to its repetitive behavior. In some cases, distance of the moving object also forms a basis for it to be considered a background, e.g. if in a scene one person is close to the camera while there is a person far away in background, in this case the nearby person is considered as foreground while the person far away is ignored due to its small size and the lack of information that it provides. [35][36]Identifying moving objects from a video sequence is a fundamental and critical task in many computer vision applications. A common approach is to perform background subtraction, which identifies moving objects from the portion of video frame that differs from the background model.

3.2. 1 Background Subtraction Algorithms

Most of the background subtraction algorithm follows a simple flow diagram as shown in Fig.2

3.2. 1. 1 Pre-processing

Frame preprocessing is the first step in the background subtraction algorithm. The purpose of this step is to prepare the modified video by removing noise and unwanted object in the frame in order to increase the amount of information gained from the frame and the sensitivity of the algorithm. Preprocessing is a process collecting a simple image processing task that change the raw input video in to a format. This can be processed by subsequent steps. Preprocessing of the video is necessary to improve the detection of moving objects by example, by spatial and temporal smoothing; snow can be removed from the video. Small moving object such as moving leave in a tree can be removed by morphological processing of the frame after the identification of the objects.[37][39]

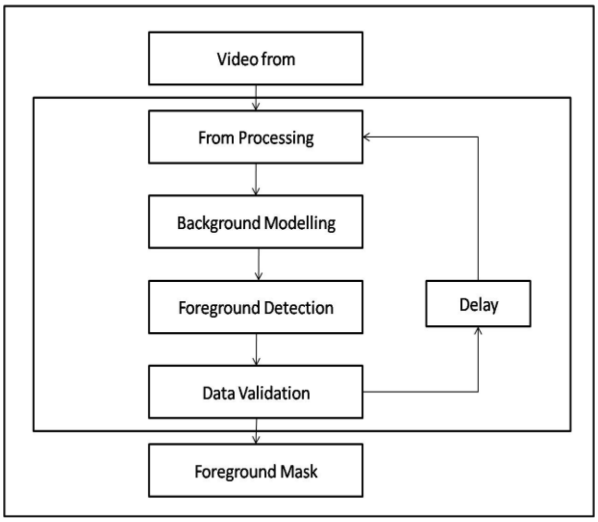


Fig 2: Flow diagram of a generic background subtraction algorithm

Another key issue in processing is the data format used by the background subtraction algorithm. Most of the algorithms handle luminance, intensity, which is one scalar value par each pixel. However color image, in either in RGB, or HSV color space, is becoming more popular in the background subtraction algorithms. There are six operations that can be performed:

1. Addition:
2. Subtraction:
3. Multi-image averaging:
4. Multi -image modal filtering**:**
5. Multi -image median filtering
6. Multi-image averaging filtering.

3.2. 1.2 Background modeling

Background modeling and subtractions core component in motion analysis. The central idea behind such module is to create a probabilistic representation of the static scene that is compared with the current input to perform subtraction. Background modeling is at the heart of any background subtraction algorithm. Background modeling uses the new video frame to calculate and update a background model. Background modeling techniques can be classified into two main categories non-recursive and recursive technique.[37][39][41]

1) **Non recursive techniques***:* A non recursive technique uses a sliding window approach for background estimation. It stores a buffer of the previous video frames, and estimate the background image based on the temporal variation of each pixel within the buffer. Non recursive technique are highly adaptive as they do not depend on the history beyond those frame stored in the buffer. On the other hand, the storage requirement can be significant if a large buffer is needed to cope with slow -moving traffic. Some of the commonly used non recursive techniques are Median Filter, Linear predictive filter, Frame Differencing.
2) **Recursive Technique***:* Recursive technique do not maintains buffer for background estimation. Instead, they recursively update a single background model based on each input frame. As a result, input frame from distant on the current background model. Compared with non-recursive techniques, recursive techniques require less storage, but any error in the background model can linger for a much longer period of time.
3) **Foreground Detection***:* Foreground detection compares the input video frame with the background model, and identifies candidate foreground pixels from the input frame. Foreground detection then identifies pixel in the video frame that cannot be adequately explained by the background model, and output them as a binary candidate foreground mask.
4) **Data Validation**: Data validation examines the candidate mask, eliminates those pixels that do not correspond to actual moving objects, and output that the final foreground mask.

3.3 Tracking

Tracking is the problem of generating interference about the motion of an object given a sequence of images. Good solution to this problem has variety of applications:

- **Motion Capture***:* If we can track a moving person accurately, than we can make an accurate record of their motion .Once we have this record, we use it to drive a rendering process; for example, we might control a cartoon character, thousand of virtual extra in a crowd scene.[10] Furthermore, we could modify the motion record to obtain slightly different motion.
- **Re cognation from motion***:* The motion of object is quite characteristic. We may be able to determine the identity of the object from its motion; we should be able to tell what it is doing.
- **Surveillance :** Knowing what objects are doing can be very useful .For example, different kinds of trucks should move in different, fixed pattern in an airport; if they do not, then something is going wrong. It could be helpful to have a computer system that can monitor activities and give warning if it detects a problem case[11].
- **Targeting:** A significant fraction of tracking literature is oriented toward (a) what to shoot, and (b) hitting it.

3.4 Optical Flow

Optical flow or optic flow is the pattern of apparent motion of objects, surfaces and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene. [45][46]The concept of optical flow was first studied in the 1940s and ultimately published by American psychologist James J. Gibson as part of his theory of affordance. Optical flow techniques such as motion detection, object segmentation, time-to-collision and focus of expansion calculations, motion compensated encoding, and stereo disparity measurement utilize this motion of the objects surfaces, and edges.

3.4.1 Estimation of the optical flow: Sequences of ordered images allow the estimation of motion as either instantaneous image velocities or discrete image displacements. It emphasizes the accuracy and density of measurements.

The optical flow methods try to calculate the motion between two image frames which are taken at times t and t + t at every voxel position. These methods are called differential since they are based on local Taylor series approximations of the image signal; that is, they use partial derivatives with respect to the spatial and temporal coordinates.

Motion estimation and video compression have developed as a major aspect of optical flow research. While the optical flow field is superficially similar to a dense motion field derived from the techniques of motion estimation, optical flow is the study of not only the determination of the optical flow field itself, but also of its use in estimating the three-dimensional nature and structure of the scene, as well as the 3D motion of objects and the observer relative to the scene.

Optical flow was used by robotics researchers in many areas such as: object detection and tracking, image dominant plane extraction, movement detection, robot navigation and visual odometry. Optical flow information has been recognized as being useful for controlling micro air vehicles.

The application of optical flow includes the problem of inferring not only the motion of the observer and objects in the scene, but also the structure of objects and the environment. Since awareness of motion and the generation of mental maps of the structure of our environment are critical components of animal (and human) vision, the conversion of this innate ability to a computer capability is similarly crucial in the field of machine vision.

IV. PROBLEM DEFINITION AND FORMULATION

4. 1 Problem definition

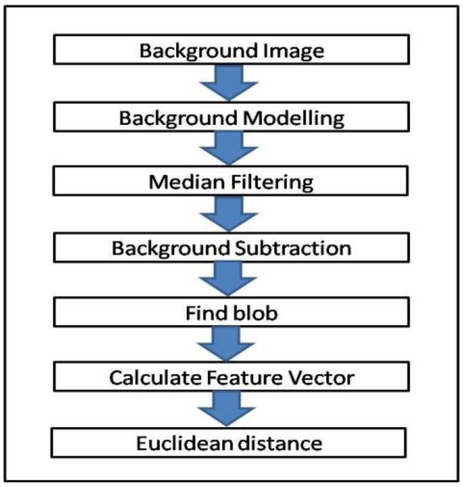
Dealing with multiple moving object in static background is a crucial challenge in object detection It is specially relevant in automatic surveillance application where accurate tracking is very important even in under crowded condition where multiple object are in motion . An efficient and robust algorithm for multiple object (human) detection from video surveillance is developed for this process; we had to perform a no of operation step wise and systematic manner.

4.2 Scope

The implementation can be used in video surveillance where the video is stable with a simple background. It can be applied to videos from a fixed camera with stability and the fluctuation is very less .The implementation can be used for many applications where the above condition is met.

4.3 Problem Formulation

We approach the problem with help of the following steps as shown in the flow chart.



V. SOLUTION APPROACHES

Our surveillance activity goes through three phases. In first phase, the target is detected in each video frame. In second phase, feature extraction is done for matching and in third phase, the detected target is tracked through a sequence of video frames.

5. 1 Assumptions

- The background is almost static. It should not change during the whole test video clip. The changes can occur due to shadow, so the video is taken in indoor environment.

- It should be free from illumination changes.
- The lens of camera should not shake during the process; it must be avoided as far as possible.
- The overlapping of two people must be avoided so that the problem of occlusion never arises.
- Moving object in the video should not be very far from camera.

5.2 Computer Algorithm:

In our algorithm we first take a suitable video having no moving object in it, so that, the background (reference) image can be extracted easily. We build an initial statistical model for a background scene that allows us to detect foreground regions even when the background scene is not completely stationary. The system updates the background model parameters adaptively to decrease the number of false positives.

Then we have to model the background image which contains the non-moving objects in a video. Obtaining a background model is done in two steps: First, the background initialization where we obtain the background image from a specific time from the video sequence. In the second step, the background maintenance is done.

A medium filter is applied afterwards to reduce noise. We then apply the Background subtraction method which is used for object detection. In this method the background objects is subtracted from the current image and thereby obtain the object, then the detected objects are converted into image blobs defined as bounding boxes representing the foreground objects so that significant features can be extracted from them. These features are for matching blobs with corresponding blobs in sequence of frames. The coherent pixels are grouped together as image blobs by seeded region growing approach. After finding all the image blobs, smaller ones are discarded. Many features can be used for matching purpose. Some significant features of blobs for matching purpose can be considered as:

- Size of blob
- Average of individual RGB components
- co-ordinate of centre of blob
- Motion vector
- Distance between the blob

We consider the size of the blob and co-ordinate of center of blob as feature for matching to be done. We then calculate the feature vector for each and every blob belonging to corresponding frame and this is to be applied to all the frames in the video.

- Take a Background image
- Model the Background image
- Apply Median filter to remove noise
- Use Background subtraction

  Image = current image - background image
- Find the blob for feature extraction
- Calculate feature vector for each blob
- Calculate the Euclidian distance between blob pair
- Find the minimum Euclidian distance

5.3 Mathematical Analysis

Tracking is performed by matching features of blobs in the current frame with the features of the blobs in the previous frame. The difference between the feature vectors of each blob in current frame with each of previous frame is calculated. We do an exhaustive matching among N blobs in the current frame with M blobs in the previous frames, so a total of NxM matching is required. As we do not have a lot of objects in the scene, this exhaustive matching is not time consuming. This difference is obtained by using Euclidian distance given by equation 1 :

$$\mathrm{Dist}(E_i,E_j)=\sqrt{\frac{1}{d}\sum_{k=1}^{d}(E_{ik}-E_{jk})^2} \qquad \text{...........}\ 1$$

Where E_i and E_j are feature vectors And d indicates dimension of the vector.

The corresponding minimum distance between two blobs feature vectors is selected and remaining are discarded. Selected blob pair is the tracked blob from the previous one to current one. This process is continued for complete video and thus tracking of multiple people is achieved.

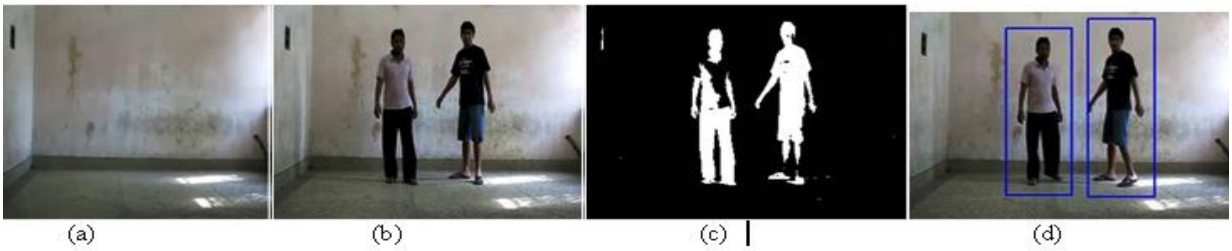


Fig 3 : A video (240x320) is captured for the simulation. A Background image is taken from the scene as shown in fig (a). At any time t a frame containing the foreground objects along with background image is taken from the video as shown in (Fig b). Foreground image (Fig.c) is calculated by subtracting current image with background using matlab image toolbox detected blob (Fig. d) has been found.

In this above algorithm, we have presented methods of segmentation of foreground object by background subtraction and tracking of multiple people in indoor environment. We selected background subtraction method, because it gives maximum number of moving pixels. We used feature based tracking, as it is faster than other methods.
There are some problems associated with this method:

- Occlusion handling problem i.e. overlapping of moving blobs has to be dealt carefully.
- Human locomotion tracking
- Lighting condition.
- shaking camera
- shadow detection
- People in shape, color and size also pose a great challenge to efficient tracking.

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

- A statistical background appearance model directs the system’s attention to the regions showing difference from the background.
- A camera model to provide a transformation from the world to the image. In conjunction with the assumption that humans move on a known ground plane, it helps transform positions between the image and the physical world and allows reasoning with invariant 3D quantities (e.g., height and shape).
- A 3D coarse human shape model to constrain the shape of an upright human. It is critical for human segmentation and tracking.
- A 3D human articulated locomotion model to help recover the locomotion modes and phases and recognize walking humans to eliminate false hypotheses formed by the static analysis.

The overview block diagram of the system is shown in Fig. 2. First, the foreground blobs are extracted by a change detection method. Human hypotheses are computed by boundary analysis and shape analysis using the knowledge provided by the human shape model and the camera model. Each hypothesis is tracked in 3D in the subsequent frames with a Kalman filter using the object’s appearance constrained by its shape. Two-dimensional positions are mapped onto the 3D ground plane and the trajectories are formed and filtered in 3D. Depth ordering can be inferred from the 3D information, which facilitates the tracking of multiple overlapping humans and occlusion analysis.

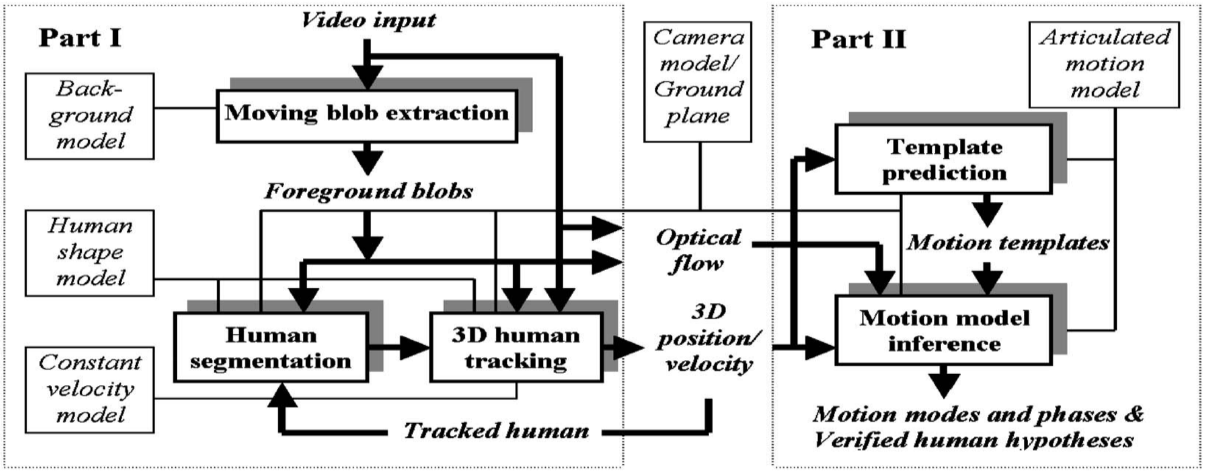


Fig. 4: The system diagram. Shaded box: program module; plain box: model; thick arrow: data flow; thin line: model association.

VI. SEGMENTATION AND TRACKING OF MULTIPLE HUMANS

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

We incorporate a statistical background model [44] 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 non-moving pixels. A single initial background frame is sufficient to start. The background model can be easily replaced with a more complex one (e.g., one with a multi-Guassian model [40] or one which can start with moving objects in the scene [15]) 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.

In contrast to the ground-level camera setup used in some of the previous work (e.g., [15], [25], etc.), we deploy the camera a few meters above the ground looking down. 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 3D feature points ($_3$ 6 points with $_3$ 2 of them out of a plane) and their corresponding image points. A linear calibration method described in [12] 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 (e.g., [22], [24]). It has also been shown in [24] that humans walking in more than one direction can provide enough information for an approximate camera calibration. Both methods have been used in our experiments.

We 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 3D quantities. 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, we only 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.

We 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. Relation between them is given in [16] by $C^{-1} = PQ^{-1}P^{T}$. An object mask M is defined by the pixels inside the ellipse. The 3D human shape model also enables geometric shadow analysis.

6.2 Segmenting Multiple Humans

We attempt to interpret the foreground blobs with the ellipsoid shape model. Human hypotheses are generated by analysing the boundary and the shape of the foreground blobs. The process is described below and shown step by step graphically in Fig. 5.

6.2.1 Locating People by Head Top Candidates

In scenes with the camera placed several meters above the ground, the head of a human is less likely to be occluded; we find that recognizing the head top on the foreground boundary is a simple and effective way to locate multiple, possibly overlapping humans.

A point can be a head top candidate if it is a peak (i.e., the highest point in the vertical direction (the direction towards the vertical vanishing point) along the boundary within a range (Fig. 5a)) defined by the average size of a human head assuming an average height. A human model of an average height is placed at each peak.

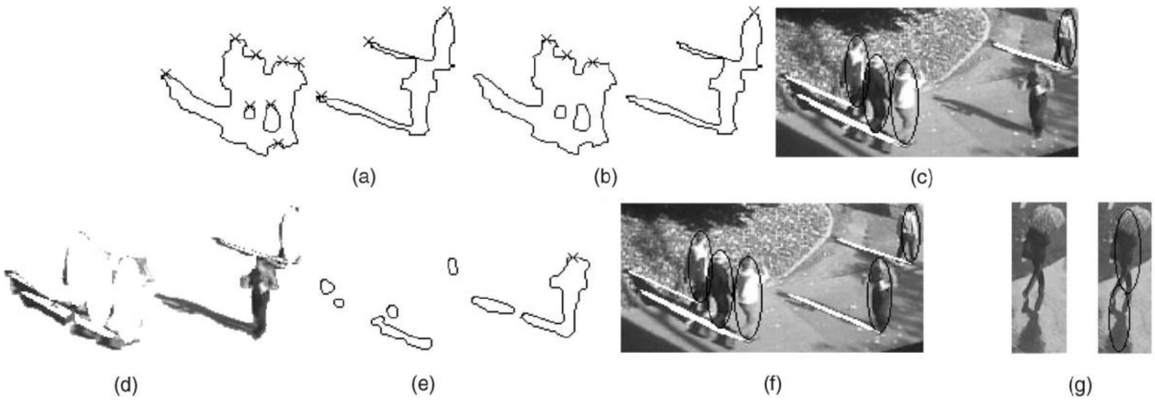


Fig. 5. The process of multihuman segmentation. (a) unscreened head top candidates; (b) screened head top candidates; (c) first four segmented people; (d) the foreground residue after first four people are segmented; (e) head top candidate after first four people are segmented; (f) the final segmentation; (g) an example of false hypothesis

Those peaks which do not have sufficient foreground pixels within the model are discarded (Fig. 5b). If a head is not overlapped with the foreground region of other objects, it is usually detected with this method (Fig. 5c).

For each head top candidate, we find its potential height by finding the first point that turns to a background pixel along the vertical direction in the range determined by the minimum and the maximum human height. We do this for all points in the head area and take the maximum value; this enables finding the height of different human postures. Having head top position and the height, an ellipsoid human hypothesis is generated.

6.2.2 Geometrical Shadow Analysis

Assuming that the sun is the only light source and its direction is known (can be computed from the knowledge of time, date, and geographical location, e.g., using [29]), the shadow of an ellipsoid on the ground, which is an ellipse, can be easily determined. Any foreground pixel which lies in the shadow ellipse and whose intensity is lower than that of the corresponding pixel in the background by a threshold T_s is classified as a shadow pixel. Most of the current shadow removal approaches are based on an assumption that the shadow pixels have the same hue as the back-ground but are of lower intensity (see [33] for a review) and ignore the shadow geometry. The color-based approaches are not expected to work well on very dark sun cast shadows, as hue computation will be highly inaccurate.

6.2.3 The Algorithm

Segmenting multiple humans is an iterative process. We denote the foreground mask after removing the existing human masks and their shadows as the foreground residue map Fr. At the beginning of the segmentation, Fr is initialized with F . The head top candidate set Hc is computed from Fr. We choose one candidate, which has the minimum depth value (closest to the camera) to form a human hypothesis. Figs. 5c and 5d show the first four segmented humans and the foreground after their masks and shadow pixels are removed. As can be seen, a large portion of the shadow pixels is removed correctly. A morphological open operation is performed on Fr to remove the isolated small residues (Fig. 5e). This process iterates until no new head candidates are found (Fig. 5f).[35][44]

This approach works well for a small number of overlapping people that do not have severe occlusion; a severely occluded object will be detected when it becomes more visible in a subsequent frame. This method is not sensitive to blob fragmentation if a large portion of the object still appears in the foreground. In our experiments, we found that this scheme tends to have a very low false alarm rate. The false alarms usually correspond to large foreground region not (directly) caused by a human. For example, when people move with their reflections, the reflections are also hypothesized as humans

6.3 Tracking Multiple Humans

Once segmented, the objects are tracked in the subsequent frames. Tracking is a loop consisting of prediction of the positions from the previous frame, search for the best match,and update of the object representation. Multiple objects are matched one by one according to their depth order. Object Representation for Tracking An elliptic shape mask (M) projected from the ellipsoid model represents the gross human shape. The shape/scale of the mask changes automatically according to the human’s position and the geometry. A texture template (T) is used to represent the appearance of a human by the rgb value of each pixel. Not every pixel inside the elliptic mask corresponds to the foreground; we also keep a foreground probability template (Fp) for each human object, which stores the probability of each pixel in the elliptic mask as foreground. It enables handling of some variations of body shape/pose.

Fig. 6b shows examples of the representation. Due to camera perspective effect, the elliptic masks of the same ellipsoid have different shape (i.e., orientations and lengths of the axes) when the human is at different locations. Therefore, a mapping is needed to align different ellipses for matching and updating. Suppose we have two ellipses e1 (u1,α1,β1,ϴ1) and e2 (u2,α2,β2,ϴ2) in their parametric forms where u, α, β and ϴ are the center, long axis, short axis, and the rotation, respectively. A mapping u’ =W(u) transforms a point u in e1 to its corresponding point u’ in e2 by aligning e1 and e2 with their centers and corresponding axes through translation, rotation, and scaling by equation 2,3 & 4.

$$u' = W(u) \qquad \text{..........................2}$$

$$= \begin{bmatrix} \cos\theta_2 & -sin\theta_2 \\ sin\theta_2 & cos\theta_2 \end{bmatrix} \begin{bmatrix} \frac{\alpha_2}{\alpha_1} & 0 \\ 0 & \frac{\beta_2}{\beta_1} \end{bmatrix} \begin{bmatrix} cos\theta_1 & sin\theta_1 \\ -sin\theta_1 & cos\theta_1 \end{bmatrix} \qquad \text{............................3}$$

$$= [u - u_1] + u_2 \qquad \text{.............................4}$$

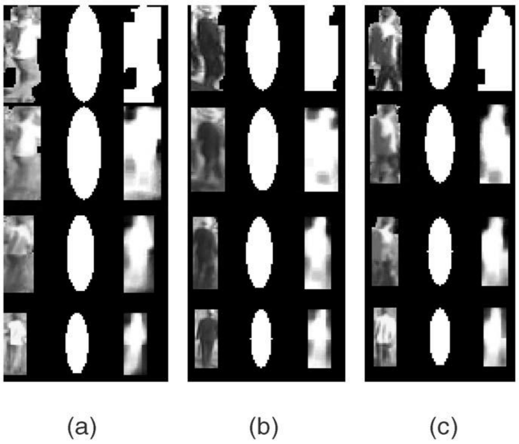


Fig. 6. Examples of object representation for tracking and its evolution: (a) texture template, (b) shape mask, and (c) foreground probability template. From top to bottom: 1st, 25th, 100th, 200th frame, respectively

6.4 Handling Occlusions

Occlusion of multiple objects has been addressed in several places in the algorithm, for example, in matching and updating. Furthermore, we compute r, the visible fraction of the object. r is defined by Nv/Ne, where Nv is the number of visible (i.e., unoccluded) foreground pixels in the elliptic mask and Ne is area, in pixel, of the elliptic mask of each object. The measurement noise n1,n2 of the Kalman filter are set proportional to 1/r. Using two thresholds To1 and To2, if To1 > r > To2, the object is said to be partially occluded. If r < To2, the object is said to be completely occluded. In case of complete occlusion, the object follows the prediction of the Kalman filer. If an object is completely occluded for a certain number of frames, it is discarded. [27][47]

VII. Conclusion & Future Work

We have presented methods of segmentation of foreground object by background subtraction and tracking of multiple people in indoor environment. We selected background subtraction method, because it gives maximum number of moving pixels. We used feature based tracking, as it is faster than other methods. Then described our methods for segmentation and tracking of multiple humans in complex situations and estimation of human locomotion models that address the problem of occlusions in the tracking process.

There are a few interesting directions to be explored in the future. A joint likelihood might be needed in segmentation and tracking of more overlapping objects. Further, using 2 cameras to construct 3D human models that would give more precise results. In future Extraction of foreground Object from dynamic scene will be emphasized along with variable light condition and different camera angle. Motion parameters and body parameters can be optimized locally to best fit the images.

References

[1] A.M. Baumberg, “Learning Deformable Models for Tracking Human Motion,” PhD thesis, Univ. of Leeds, 1995.

[2] G.A. Bekey, “Walking,” The Handbook of Brain Theory and Neural Networks, M.A. Arbib, ed., MIT press, 1995.

[3] M. Brand, “Shadow Puppetry,” Proc. Int’l Conf. Computer Vision, vol. 2, pp. 1237-1244, 1999.

[4] C. Bregler, “Learning and Recognizing Human Dynamics in Video Sequences,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 568-574, 1997.

[5] A.F. Bobick and J.W. Davis, “The Recognition of Human Movement Using Temporal Templates,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 3, Mar. 2001.

[6] Character Studio: Software Package, http://www.discreet.com/ products/cs/, 2002.

[7] I. Cohen and G. Medioni, “Detecting and Tracking Moving Objects for Video Surveillance,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 319-325, 1999.

[8] R. Cutler and L.S. Davis, “Robust Real-Time Periodic Motion Detection, Analysis, and Applications,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug. 2000.

[9] J. Deutscher, A. Davison, and I. Reid, “Automatic Partitioning of High Dimensional Search Spaces

Associated with Articulated Body Motion Capture," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 669-676, 2001.
[10] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," Proc. IEEE Int'l Conf. Computer Vision, pp. 726-733, 2003.
[11] A.M. Elgammal and L.S. Davis, "Probabilistic Framework for Segmenting People under Occlusion," Proc. Int'l Conf. Computer Vision, vol. 1, pp. 145-152, 2001.
[12] D. Forsyth and J. Ponce, Computer Vision: A Modern Approach. Prentice-Hall, 2001.
[13] S. Hongeng and R. Nevatia, "Multi-Agent Event Recognition," Proc. Int'l Conf. Computer Vision, vol. 2, pp. 84-91, 2001.
[14] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4S: A Real-Time System for Detecting and Tracking People in 2 1/2 D," Proc. European Conf. Computer Vision, pp. 962-968, 1998.
[15] S. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities," IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug. 2000.
[16] R. Hartley and A. Zisserman, Multi View Geometry. Cambridge Press, 2000.
[17] M. Isard and A. Blake, "Condensation-Conditional Density Propagation for Visual Tracking," Int'l J. Computer Vision, vol. 29, no. 1, pp. 5-28, 1998.
[18] M. Isard and J. MacCormick, "BraMBLe: A Bayesian Multiple- Blob Tracker," Proc. Int'l Conf. Computer Vision, vol. 2, pp. 34-41, 2001.
[19] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," J. Basic Eng., vol. 82, pp. 35-45, 1960.
[20] P. Kornprobst and G. Medioni, "Tracking Segmented Objects Using Tensor Voting," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 118-125, 2000.
[21] N. Kra hnstover, 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.
[22] D. Liebowitz, A. Criminisi, and A. Zisserman, "Creating Architectural Models from Images," Proc. EUROGRAPH Conf., vol. 18, pp. 39-50, 1999.
[23] 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.
[24] 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.
[25] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking Groups of People," Computer Vision and Image Understanding, vol. 80, no. 1, pp. 42-56, 2000.
[26] T.B. Moeslund and E. Granum, "A Survey of Computer Vision- Based Human Motion Capture," Computer Vision and Image Understanding, vol. 81, pp. 231-268, 2001.
[27] G . Mori and J. Malik, "Estimating Human Body Configurations Using Shape Context Matching," Proc. European Conf. Computer Vision, pp. 666-681, 2002.
[28] R. Mu rry, Z.X. Li, and S. Sastry, A Mathematical Introduction to Robotic Manipulation. CRC Press, 1994.
[29]NOVAS—NavalObservatory Vector Astrometry Subroutines,
http://aa.usno.navy.mil/software/novas/novas_info.html, 2003.
[30] Data Set Provided by IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS2001), 2001.
[31] S. Pingali and J. Segen, "Performance Evaluation of People Tracking Systems," Proc. Third IEEE Workshop Applications of Computer Vision, pp. 33-38, 1996.
[32] P.J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K.W. Bowyer, "The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm," Proc. Int'l Conf. Pattern Recognition, pp. 385- 388, 2002.
[33] A. Prati, R. Cucchiara, I. Mikic, and M.M. Trivedi, "Analysis and Detection of Shadows in Video Streams: A Comparative Evaluation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 571-576, 2001.
[34] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Slected Applications in Speech Recognition," Proc. IEEE, vol.77, no. 2, 1989.
[35] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," CVGIP: Image Understanding, vol. 59, no. 1, pp. 94-115, 1994.
[36] 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.
[37] 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.
[38] N.T. Siebel and S. Maybank, "Fusion of Multiple Tracking Algorithm for Robust People Tracking," Proc. European Conf. Computer Vision, pp. 373-387, 2002.
[39] Y. Song, X. Feng, and P. Perona, "Towards Detection of Human Motion," Proc. IEEE Conf. Computer

Vision and Pattern Recognition, pp. 810-817, 2000.
[40] 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.
[41] H. Tao, H.S. Sawhney, and R. Kumar, "A Sampling Algorithm for Tracking Multiple Objects," Proc. IEEE Workshop Vision Algorithms, 1999.
[42] 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.
[43] A.M. Tekalp, Digitial 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.
[45] 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.
[46] 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.
[47] T. Zhao, "Model-Based Segmentation and Tracking of Multiple Humans in Complex Situations," PhD thesis, Univ. of Southern California, Los Angeles, 2003.

AUTHORS:

Shalini Agarwal: I am Shalini Agarwal student of M.Tech (Computer Science) 2nd year from Banasthali Vidhyapeeth, Rajasthan. I have completed B.Tech (Computer Science and Engineering) in 2009 at B.S.A.C.E.T., Mathura (U.P.). My area of interest is Pattern Recognition & Image Processing, Data Mining.

Shaili Mishra: I am Shaili Mishra student of M.Tech (Computer Science) 2nd year from Banasthali Vidhyapeeth , Rajasthan. I have completed MCA in 2009 at S.R.M.C.E.M; Lucknow (U.P.).My area of interest is Pattern Recognition & Image Processing, Algorithms.