Tracking and Object Classification for Automated Surveillance

Omar Javed and Mubarak Shah

Computer Vision Lab, University of Central Florida, 4000 Central Florida Blvd, Orlando, Florida 32816, USA
{ojaved,shah}@cs.ucf.edu

Abstract. In this paper we discuss the issues that need to be resolved before fully automated outdoor surveillance systems can be developed, and present solutions to some of these problems. Any outdoor surveillance system must be able to track objects moving in its field of view, classify these objects and detect some of their activities. We have developed a method to track and classify these objects in realistic scenarios. Object tracking in a single camera is performed using background subtraction, followed by region correspondence. This takes into account multiple cues including velocities, sizes and distances of bounding boxes. Objects can be classified based on the type of their motion. This property may be used to label objects as a single person, vehicle or group of persons. Our proposed method to classify objects is based upon detecting recurrent motion for each tracked object. We develop a specific feature vector called a 'Recurrent Motion Image' (RMI) to calculate repeated motion of objects. Different types of objects yield very different RMI's and therefore can easily be classified into different categories on the basis of their RMI. The proposed approach is very efficient both in terms of computational and space criteria. RMI's are further used to detect carried objects. We present results on a large number of real world sequences including the PETS 2001 sequences. Our surveillance system works in real time at approximately 15Hz for 320x240 resolution color images on a 1.7 GHz pentium-4 PC.

1 Introduction

Automatic detection and recognition of objects is of prime importance for security systems and video surveillance applications. Automated video surveillance addresses real time observation of people and vehicles within a busy environment. Outdoor surveillance systems must be able to detect and track objects moving in its field of view, classify these objects and detect some of their activities. In this paper first we discuss the issues that any good surveillance system needs to cope with. These include problems faced in detection of objects, lighting conditions, shadows, the different types of occlusions that occur in the scene and entries/exits of objects.

Existing surveillance systems can be classified into categories according to, the environment they are primarily designed to observe i.e. indoor, outdoor or

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353, pp. 343–357, 2002.

airborne, the number of sensors i.e single camera vs. multiple cameras etc. A large number of surveillance systems have been proposed in recent years. For instance, PFinder [2] uses a unimodal background model to locate interesting objects. It tracks the full body of a person though it assumes that only a single person is present in the scene . In the approach proposed by Stauffer and Grimson [1] an adaptive multi-modal background subtraction method that can deal with slow changes in illumination, repeated motion from background clutter and long term scene changes is employed. After background subtraction the detected objects are tracked using a multiple hypothesis tracker. Common patterns of activities are statistically learned over time and unusual activities are detected. Ricquebourg and Bouthemy [4] proposed tracking people by exploiting spatio-temporal slices. Their detection scheme involves the combined use of intensity, temporal differences between three successive images and of comparison of the current image to a background reference image which is reconstructed and updated online. They track the apparent contours of moving articulated structures by using spatio-temporal slices from the image sequence volume XYT. A simple classification between persons and vehicles is also performed by using the recovered spatio-temporal signatures of each object. W4 [3] uses dynamic appearance models to track people. Single person and groups are distinguished using projection histograms. Each person in a group is tracked by tracking the head of that person. A recursive convex hull algorithm is used to find body part locations for single person. Symmetry and periodicity analysis of each silhouette is used to determine if a person is carrying an object.

In our view, tracking is the most important but error prone component of a surveillance system. Robust classification and activity detection of objects is impossible if tracking is inaccurate. In this paper we formulate object tracking as a region correspondence problem, given background subtraction results from [1]. A number of approaches have been proposed to solve the point correspondence problem with deterministic algorithms in particular see [6] [7] [8]. These algorithms are conceptually simple, have few parameters and usually address point correspondence for a dense set of points. However, due to noisy background subtraction, change in the size of regions, occlusion and entry/exit of objects, traditional point correspondence methods cannot be directly applied to the human tracking problem. We describe the problems encountered in establishing correct correspondence, and present a solution based on linear velocity, size and distance constraints.

Most of the surveillance systems do not tackle the problems in tracking caused by shadows. Azerbayjani et al.[2] proposed background subtraction in normalized color space i.e. For a yuv image, they divide the u and v components by y to achieve illumination invariance. Horprasert et el. [9] proposed a color model that separates the brightness from the chromaticity component in rgb space. These approaches can only deal with light shadows. The illumination variation in strong shadows can not be handled by such normalization procedures. Rosin and Ellis [10] proposed an approach which first marks potential shadow areas using change in color with respect to background i.e. shadows are always darker

than background. Then they use a region growing algorithm which uses a growing criterion based on the fixed attenuation of photometric gain over the shadow region compared to background. They assume that the gain will be approximately constant over the shadow region. However we have observed that shadow regions can have a variety of intensities depending upon the background. We propose a shadow detection approach based on similarity of background and shadow regions.

Once the tracking of object is achieved, we are interested in determining its type. Ideally, object classification can be attempted by using shape information from a single image. However, the work on object recognition in the past 30 years has demonstrated that object recognition or classification from a single image is a highly complex task. We believe that motion based classification reduces the reliance on the spatial primitives of the objects and offers a robust but computationally inexpensive way to perform classification. We present a solution to this problem using temporal templates. Temporal templates are used for classification of moving objects. A temporal template is a static vector image in which the value at each point is a function of motion properties at the corresponding spatial location in the image sequence. Motion History and Motion Energy images are examples of temporal templates, proposed by Bobick and Davis [5]. Motion History image is a binary image with a value of one at every pixel where motion occurred. In Motion History image pixel intensity is a function of temporal history i.e. pixels where motion occurred recently will have higher values as compared to other pixels. These images were used for activity detection. We have defined a specific Recurrent Motion template to detect *repeated* motion. Different types of objects yield very different Recurrent Motion Images (RMI's) and therefore can easily be classified into different categories on the basis of their RMI. We have used the RMIs for object classification and also for carried object detection.

The paper is organized as follows: Section 2 discusses the object detection and tracking problem. First the general problems faced by surveillance systems in outdoor environments are described. Then our solutions to the shadow detection and motion correspondence problems are presented. Section 3 focuses on the methods to classify objects into single person, groups of persons and vehicles. An efficient method to detect carried objects is also described here. Finally results are presented in Section 4.

2 Tracking in a Single Camera

Our surveillance approach is based upon extracting objects in the form of regions from the scene using a background subtraction method, tracking these objects using region correspondence, classifying these objects into people, groups of people and vehicles and then finally performing simple activity detection.

2.1 Important Problems in Realistic Scenarios

Detection and tracking of objects in a static camera is a nontrivial task. A number of problems including change in illumination, shadows, occlusion etc arise in realistic environments which need to be dealt with, by the surveillance systems.

Object Detection. The first problem for automated surveillance is the detection of interesting objects in visible range of the video camera. The objects can be persons, vehicles, animals. In the rest of the paper, we have used the term 'object' to denote any interesting object. The term 'scene structure' is used to denote inanimate objects in the scene for example trees, flags e.t.c. Almost all outdoor surveillance systems employ some variant of background subtraction methods to extract objects from the scene. However the background subtraction methods can't deal with the following problems.

- Quick changes in lighting conditions completely change the color characteristics of the background. Almost all real time background subtraction methods can't model quick and large illumination variations. Thus surveillance under partially cloudy days will fail.
- Uninteresting moving objects. For example flags waving or winds blowing through trees for short burst of time. Reflection of moving objects from shiny or wet surfaces also causes problems.
- Shadows. Background subtraction methods fail to distinguish between an object and its shadow. Shadows can be of two types 1) self shadow and 2) cast shadow. The self-shadow is the part of the object, which is not illuminated by direct light. The cast shadow is the area in the background projected by the object in the direction of light rays. In outdoor images cast-shadows are major problems in acquiring accurate silhouettes. Cast Shadows make accurate silhouette analysis impossible, that is separate objects can appear to be joined together due to shadows. Inaccurate silhouettes also cause problem during classification of objects. Note that any shadow detection and removal scheme should only remove cast shadows since removal of self-shadows will result in incomplete silhouettes.

We believe that the above mentioned problems must be solved before robust object detection in real scenes is possible.

Tracking under Occlusion. The goal of tracking is to establish correspondence between objects across frames. Occlusion occurs when an object is not visible in an image because some other object/structure is blocking its view. Tracking objects under occlusion is difficult because accurate position and velocity of an occluded object can't be determined. Different cases of occlusion are described in the following

- Inter-object occlusion occurs when one object blocks the view of other objects in the field of view of the camera. The background subtraction method gives a single region for occluding objects. If two initially non-occluding objects cause occlusion then this condition can be easily detected. However if objects enter the scene occluding each other then it is difficult to determine if inter-object occlusion is occurring. The problem is to identify that the foreground region contains multiple objects and to determine the location of each object in the region. Since people usually move in groups, which results in frequent inter-object occlusion so detecting and resolving inter-object occlusion is important for surveillance applications.
- Occlusion of objects due to thin scene structures like poles or trees causes an object to break into two regions. Thus more than one extracted region can belong to the same object in such a scenario. The problem is compounded if multiple objects are occluded simultaneously by such a structure.
- Occlusion of objects due to large structures causes the objects to disappear completely for a certain amount of time, that is there is no foreground region representing such objects. For example, a person walks behind a building, or a person enters a car. A decision has to be made whether to wait for reappearance of the objects, or determine that the object has exited the scene.

Exits and Entries of Objects from the Scene. Entry is defined as an object entering the field of view of the camera. Entries and exits are easy to detect if (exiting and entering) objects are separate in the camera view. However, detecting an entry and an exit of two (or more objects) at the same place and at the same time is difficult . If one person enters the scene at a certain position while another person leaves from the same position at the same time then this scenario needs to be distinguished from the situation in which person moves right near the exit and then start moving in the direction he came from.

We present an approach that attempts to solve some of the above mentioned problems.

2.2 Background Subtraction

We use the adaptive background subtraction method proposed by Stauffer and Grimson [1]. In the method, a mixture of K Gaussian distributions adaptively models each pixel intensity. The distributions are evaluated to determine which are more likely to result from a background process. The method reliably deals with long term changes in lighting conditions and scene changes. However it can't deal with sudden movements of uninteresting objects like flags waving or winds blowing through trees for short burst of time. A sudden lighting change will cause the complete frame to be denoted as foreground, if such a condition arises then the algorithm is reinitialized.

The background subtraction method gives foreground pixels in each new frame. The foreground pixels are then segmented into regions using the connected components algorithm.

2.3 Shadow Removal

There are a number of cues that provide information regarding the presence of a shadow. For instance, pixels in the shadow regions are darker than those in the reference background. Also shadows retain some texture and color information of the underlying surface under general viewing conditions.

Our method uses these cues in a hierarchical fashion. First, all foreground regions in the image that are darker than the reference images are extracted. Each potential shadow region can consist of sub regions of cast shadows, self-shadows and parts of the object darker than reference image. We perform color segmentation on each potential shadow region.

The goal is to divide the potential shadow region into sub regions, where each sub region belongs only to one of the three (cast shadow, self shadow, dark object) regions. Note that cast shadow and self-shadow regions will have different colors if the background and object are of different colors, which is usually the case.

A fast color segmentation algorithm is required due to the constraint of real time performance by surveillance systems. Also the regions to be segmented constitute only a subset of the total foreground regions. Thus the total area to be segmented is a small percentage of the total image area.

We use a K-means approximation of the EM algorithm to perform color segmentation. Each pixel value x in a potential shadow region is checked against existing K Gaussian distributions until a match is found. A match is defined if Mahalanobis distance of the pixel is less than a certain threshold. If the pixel matches a certain distribution the mean of that distribution is updated as follows

$$\mu_{n+1} = \mu_n + \frac{1}{n+1}(x_{n+1} - \mu_n) \tag{1}$$

Where x is the color of the pixel and μ_n is the mean of the Gaussian before the $n+1$th pixel was added to the distribution. Covariance of each distribution is fixed. If none of the distributions match the pixel then a new distribution is added with its mean equal to x. The process continues until all pixels are assigned some distribution.

A connected component analysis is performed so that spatially disconnected segments are divided into multiple connected segments. Afterwards region merging is performed in which smaller segments are merged with the neighboring segment with the largest area. Once the segmentation is complete we assume that each shadow candidate segment belongs to only one of three types of potential shadow regions i.e. cast shadow, self-shadow, dark object. To distinguish the cast shadow segment from the other two we use the property of shadow that it retains some texture of the underlying surface. Note that we already have the background image representation in the form of Gaussian distributions modelling the background processes. Now for each pixel location in the shadow candidate segment a comparison between the background and the current image needs to be made. However, illumination would be very different in the background and

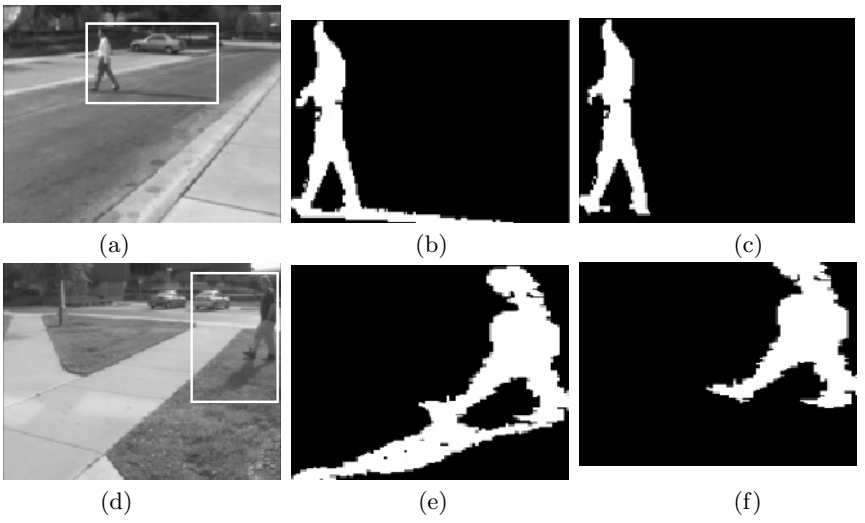

Fig. 1. Results of shadow removal. (a) and (d) show the calculated bounding box with shadow removal. (b) and (e) show the background subtraction results (zoomed). (c) and (f) show the silhouette after the segments belonging to the cast shadows have been removed.

the shadow candidate segment. Edges and gradients are good illumination invariant features. A discussion on this can be found in [11] and [12]. we use the gradient direction

$$\theta = \arctan \frac{f_y}{f_x} \tag{2}$$

for comparison of the candidate segment with the respective background region. where f_x and f_y are the horizontal and vertical derivatives respectively. The idea here is that if a region undergoes a change in illumination both f_x and f_y will undergo a change in value but their ratio will remain approximately the same. Thus for each shadow candidate segment and its respective background the gradient's are correlated. If the correlation results in more than .75 match then the candidate segment is considered a cast shadow, and is removed from the foreground region. Otherwise, the candidate segment is a self-shadow or it is the part of the silhouette. The same process is repeated for all candidate segments.

2.4 Motion Correspondence

In our approach, the goal of tracking is to establish motion correspondence between regions that are moving in a 2D space, that is essentially the projection of a 3D world. We assume the regions can enter and exit the space and they can also get occluded by other regions.

For motion correspondence, an extension of the point correspondence paradigm is used. Regions, as compared to points, carry extra information like

shape and size. This information can be used to further constrain the correspondences.

Note that uninteresting objects like trees and flags can also show up as foreground regions for short periods of time. To prevent these objects from affecting the tracking results, we establish a minimum initial observation parameter O_{min}. If an object disappears in less than O_{min} frames then it is considered a false detection. Also, the objects can disappear in the middle of the frame, for example, a person entering a building. We introduce the maximum missed observation parameter M_{max} to capture this situation. The track of a region is terminated if it is not observed in M_{max} frames.

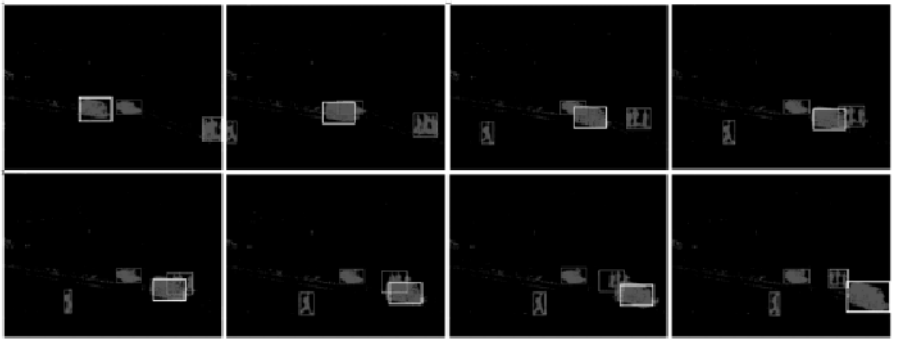

Fig. 2. Complicated occlusion example; Occlusion between 2 cars and a group of people was handled successfully. PETS Data Set 1,Testing, Camera 1 sequence. Frames 795-940

Each region is defined by the 2D coordinates of the centroid, X, the bounding box, B and the size, S. The regions, for which correspondence has been established, have an associated velocity, V, and predicted change in size, ∇S . In frame t of a sequence, there are N regions with centroids X_i^t (where $1 \leq i \leq N$) whose correspondences to previous frame are unknown. There are M regions with centroids X_L^{t-1} (where L is the label) in frame t-1 whose correspondences have been established with the previous frames. The number of regions in frame t might be less than the number of regions in frame $t-1$ due to exits or occlusion and it can be larger due to entries. The task is to establish correspondence between regions in frame t and frame $t-1$, and to determine exits and entries in these frames.

The minimum cost criteria is used to establish correspondence. The cost function between two regions is defined as

$$C_{Li} = \rho \| (X_L^{t-1} + V_L^{t-1}) - X_i^t \| + (1-\rho) \mid (\nabla S_L^{t-1} + S_L^{t-1} - S_i^t) \mid \qquad (3)$$

where

L $\in$ Labels of regions in frame $t-1$

i is index of non-corresponded region in frame t and $1 \leq i \leq N$

ρ is the weight parameter determining the percentage of cost due to change in size, and change in velocity.

The cost is calculated for all (L, i) pairs. Correspondence is established between the pair (L', i') that gives the lowest cost, with the cost being less than a threshold. The velocity and predicted size of region L' are updated using linear prediction models.

Next, all region pairs containing L' or i' are removed from consideration and the correspondence is established between the pair that gives the lowest cost among the rest of the pairs. The process continues till no pairs are left or the minimum cost rises above the threshold. The following two cases may happen at the end of the minimum cost correspondence procedure.

- Correspondences have been found between all regions in frames $t-1$ and t.
- There might be regions in frame $t-1$, which have not been corresponded to in frame t due to occlusion or due to exits from field of view of the camera. There may be regions in frame t, which have not been corresponded to regions in frame $t-1$, because they just entered the frame and no corresponding region in the previous frame exists. First we deal with frame $t-1$. Suppose a region represented by centroid X_L^{t-1} could not be corresponded to any region in frame t. A check for exit of X_L^{t-1} from the field of view of camera is done. If the position plus predicted velocity of that region is outside the frame boundary then it is determined to have exited the frame. If this is not the case, then a check for occlusion is made. If bounding box of the centroid X_L^{t-1} i.e. B_L^{t-1} overlaps the bounding box of another region B_J^t then L is marked as an occluded region. Similarly all the hitherto non corresponded regions in frame $t-1$ overlapping B_J^t are marked as occluding each other. Note that all these regions have merged in a single region J in frame t. Now we need to update the parameters of the occluded regions. For occluded region L , $V_L^t = V_L^{t-1}$. Also if $X_L^{t-1} + V_L^{t-1}$ is within B_J^t then, $X_L^t = X_L^{t-1} + V_L^{t-1}$. Otherwise X_L^t is the point on B_J^t nearest to X_L^{t-1}. This check is to constrain the occluded centroids within J. The non-corresponded regions in frame t are set to be entries, their initial velocity and change in size are set to zero.

3 Object Classification

Our goal is to classify each moving object visible in the input video as a single person, a group of persons or a vehicle. Here we present a view based technique for classification. The first step to achieve this goal requires the detection and correspondence of each object. The tracking algorithm provides the bounding box, centroid and correspondence of each object over the frames. Then we attempt to classify objects by detecting repetitive changes in shape of the objects. In most cases the whole object is also moving in addition to local changes in shape, e.g person walking. So we need to compensate for the translation and change in scale of the object over time to detect the local change. Translation is compensated by aligning the object in subsequent frames along its centroid.

For compensation of scale, the object is scaled equally in horizontal and vertical directions such that its vertical length i.e. projected height, is equal to its vertical length at the first observation. Here, we make the assumption that the only cause of change in the projected height of an object is the variation in the object's distance from the camera. Once the objects are aligned the Recurrent Motion Image is used to determine the areas of silhouette undergoing repeated change.

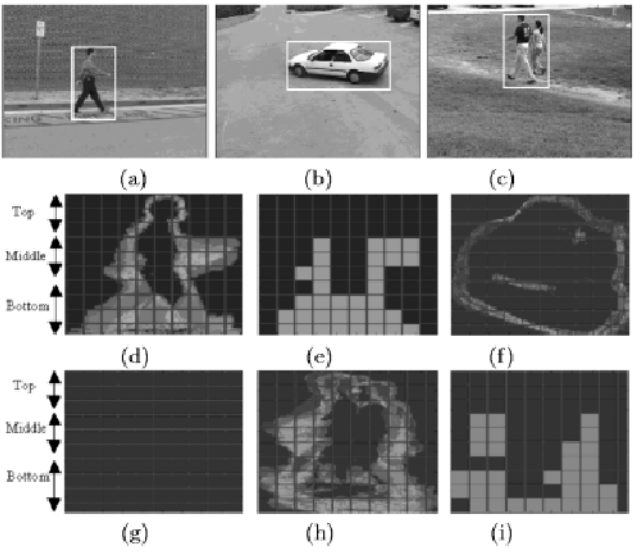

Fig. 3. (a)Single person walking (b) A moving Car (c) Two people walking (d)RMI of the single person's silhouette. (e) Blocks with average recurrence$> T$ in single person's RMI. RMI computed for 2 second intervals and $T = 2$, since we expect that a person will take atleast two strides in that time. (f) RMI of the car. (d) Average recurrence$> T$ for car RMI. (e) RMI of a group of 3 Persons. (f) Average recurrence$> T$.

3.1 Recurrent Motion Image (RMI)

Let $I_a(x, y, t)$ be a binary silhouette image sequence for an object $'a'$, obtained using background subtraction, that is translation and scale compensated. Let $D_a(x, y, t)$ be a binary image indicating areas of motion (shape change) for object $'a'$ between frame t and $t - 1$. We use the exclusive-or operator to generate the difference image $D_a(x, y, t)$ i.e.

$$D_a(x, y, t) = I_a(x, y, t - 1) \oplus I_a(x, y, t) \tag{4}$$

We define the Recurrent Motion Image (RMI), for the purpose of classification, as

$$RMI_a = \sum_{k=0}^{\tau} D_a(x, y, t-k) \tag{5}$$

In the above equation, the duration τ should be large enough to encompass the temporal extent of movement. RMI will have high values at those pixels at which motion occurred repeatedly and low values at pixels where there was little or no motion. Note that I_a is obtained using robust background subtraction, therefore RMI is not affected by small changes in lighting or repetitive motion from background clutter. The definition of RMI implies that no history of previous images needs to be stored nor manipulated, making the computation both fast and space efficient. Precisely it will take $2n\tau$ operations to compute RMIa from I_a of size n pixels.

3.2 Classification of RMI's

Once an RMI is obtained we need to classify it. A feature vector is computed by partitioning the image into N equal sized blocks and computing the average recurrence for each block. Each block belongs to top, middle or bottom section of the silhouette (see fig 3). The partitioning reduces computation and averaging reduces the effect of noise. If there are blocks in the middle and bottom sections with average recurrence value greater than a threshold T, then the object is undergoing repetitive change. Thus, it is classified as a single person or a group. If there is no recurring motion then the object is a vehicle. T depends on duration of the sequence.

The next step is to distinguish between a single person and group. Two different strategies are used to achieve this

- Shape cues are used to detect the number of heads in a silhouette. Peak points detected on the boundary of the silhouette are marked as head candidates. Head candidates which are not near significant vertical projection histogram peaks are removed. This method works well if there is some separation between heads of the people in the group.
- If the people are very close to each other then the normalized area of recurrence response in the top section of RMI is always greater than the normalized area of recurrence response in a single person due to the presence of multiple heads. Area of recurrence response in a particular section is defined as the number of blocks with average recurrence greater than one in that section. This response is thresholded to determine the presence of multiple persons.

We classify a silhouette as a group if one of the above given criteria is satisfied.

3.3 Carried Object Detection

Human Silhouettes are nearly symmetrical about the body axis while in the upright position. During walking or running, parts of arms or legs do violate the symmetry. However a sub region of silhouette not undergoing recurring local motion while consistently violating symmetry will usually be a carried object. RMI's combined with the symmetry constraint can efficiently be used to detect carried objects.

The symmetry axis of silhouettes needs to be calculated for symmetry analysis. We have observed that neither centroids nor major axis (calculated using PCA) of the silhouettes are good candidates for symmetry axis since presence of large carried or pulled objects distorts the shape of silhouette and hence of the axis. A vertical line drawn from the aforementioned head point (section 3.2) to the bottom of the object is a better candidate for symmetry axis. Since its calculation is largely unaffected by the carried objects (unless the object is carried over the head). Symmetry analysis for the vertical axis is done using a simple method. Let p_l^h,p_r^h be two points on the silhouette boundary on the horizontal line h. Let l_l^h and l_r^h be the distance of p_l^h and p_r^h respectively from the symmetry axis. A point p_x^h with distance l_x^h from the axis is classified as

$$p_x^h = \begin{cases} Nonsymmeteric & \text{if } l_x^h > min\{l_l^h, l_r^h\} + \epsilon \\ Symmetric & \text{otherwise} \end{cases} \tag{6}$$

After symmetry analysis, the RMI of the silhouette is used to determine if the non-symmetric pixels are showing recurrent local motion. All the non-symmetric pixels which do not exhibit recurrent motion are grouped into a region and marked as a carried object, as shown in fig 4.

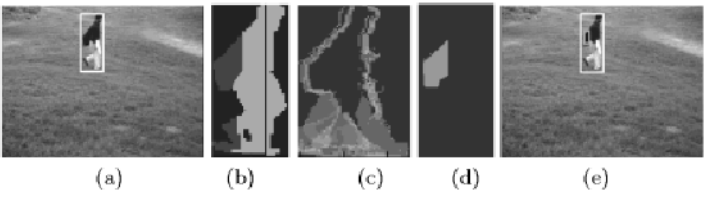

Fig. 4. (a) Object with bounding box (b)Silhouette of object (zoomed). The dark regions are the non-symmetric pixels (c) RMI of the object (d) Non-symmetric and Non recurring Pixels (e) Carried Object Detected

4 Results

The tracking and classification approach described in previous sections was applied to a variety of video sequences. The results are available on the web at,

http://www.cs.ucf.edu/∼vision/projects/Knight/Results.html. The algorithm was also applied on sequences provided for the purpose of performance evaluation of tracking in the "Second IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2001" . Specifically the video sequences (dataset 1 test sequences 1 and 2) were used. The *PETS* sequences were jpeg compressed and therefore noisy. Each sequence consists of 2688 frames. The sequences contain persons, groups or people walking together and vehicles. For each sequence, the tracking program detects moving object in the scene and calculates bounding box, centroid and correspondence of each object over the frames. The tracking algorithm successfully handled occlusions between people, between vehicles and between people and vehicles. Entry of a group of people was detected as a single entry, However, as soon as one person separated from the group he was tracked separately. One limiting assumption of the tracker is that it can't deal with division of single connected component in two large components during occlusion. It assumes that occlusion is over and updates the predicted position with wrong predicted velocities and sizes. However this region division rarely happens in cases when object are directly viewable.

Table 1. People/Vehicle classification results. Recurring Motion was detected in all RMI's for groups and single persons but not in vehicle RMI's.

Type of Object	No. of Instances	Recurrent Motion Detected	No. of Instances Correctly Classified
Single Person/Groups	23	Yes	23
Vehicles	9	NO	9

The shadow removal algorithm only kicks in if at least 30 % of the silhouette is classified as a dark region i.e. potential cast or self shadow. The PETS sequences didn't contain significant shadows. For sequences with significant shadows, the shadow removal algorithm was able to remove shadows from an object in approximately 70% of the frames in which that object was visible. In 25% of frames it didn't remove the shadow and in 5% of the frames segments belonging to an object were removed. Almost all the errors in the shadow removal results happened in the frames in which the objects had large self shadows. Error was caused by the failure of segmentation procedure to divide cast shadows and self shadow in different regions.

The RMI is calculated for each object after it has completely entered the frame. The RMI was calculated for a two second time duration. A person was expected to take at least two strides in this time. The number of frames over which the RMI was calculated was $\tau = framerate \times timeduration$. In addition to the PETS sequences, the RMI was tested on other videos taken in different conditions. Some were taken with a high angle of depression and some parallel to the ground plane. The relative size of the objects with respect to the image varied widely. The algorithm works for any silhouette greater than 50×30 pixels. Table

Fig. 5. Carried Object Detection Results.

1 gives the classification results distinguishing between persons and vehicles. Accurate classification results were obtained even in presence of noisy silhouettes. Table 2 gives the classification results distinguishing between groups and single persons. Groups of persons couldn't be distinguished from a single person if one person in the group was occluding a large area of the other persons. The carried objects are detected accurately if they satisfy the non-symmetry and non-recurrence conditions.

Table 2. Person/Group classification results

Type of Object	No. of Instances	No. of Instances Correctly Classified
Single Person	15	15
Groups	8	7

The Surveillance system has been implemented in C++ and processes 12-15 frames a second for a frame size of 320×240 pixels. The system runs on a 1.7 GHz pentium 4 machine. The System, with an added activity detection component, is currently being used by the Orlando Police Department to aid in surveillance of the Orlando (Florida, USA) downtown area. For further information please visit http://www.cs.ucf.edu/∼ vision/projects/Knight/Knight.html

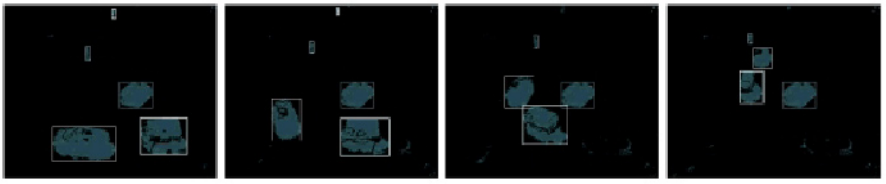

Fig. 6. Tracking results on PETS Dataset 1,test, camera 2 sequence (frames 2440-2580).

5 Conclusion

In this paper we identified important issues that need to be resolved before fully automated outdoor surveillance systems can be developed. We also present solutions to the problems of shadows, handling spurious objects, classification of objects based on recurrent motion and carried object detection. Future work includes exploring better segmentation methods to improve shadow detection.

References

1. Stauffer C. and Grimson , *"Learning Patterns of Activity using Real Time Tracking"* IEEE PAMI, Vol .22, No. 8 , August 2000, pp 747-767
2. C. Wren, A Azarbayejani, T. Darrel and A. Pentland, *"PFinder, Real Time Tracking of the Human Body"*, IEEE PAMI, vol 19, no. 7 july 1997.
3. I. Haritaoglu, D. Harwood and L. Davis, *"W4: Real time Surveillance of People and Their Activities"* IEEE PAMI Vol .22, No. 8 , August 2000.
4. Y. Ricquebourg and P. Bouthemy, *"Real Time Tracking of Moving Persons by Exploiting Spatiotemporal Image Slices "* IEEE PAMI Vol .22, No. 8 , August 2000.
5. A .Bobick and J. Davis, *"The Recognition of Human Movements Using temporal Templates"* IEEE PAMI, Vol 23, No. 3, March 2001.
6. I. K Sethi, R. Jain, *"Finding trajectories of feature points in monocular image sequences"* IEEE PAMI, Jan 1987.
7. K. Rangarajan, M. Shah, , *"Establishing motion correspondence"* CVGIP, July 1991.
8. C. J. Veenman, M.J.T. Reinders, E. Baker, *""Resolving motion correspondence for densely moving points"*IEEE PAMI Jan 2000.
9. T. Horprasert, D. Harwood and L. Davis , *"A Statistical Approach for Real Time Robust Background Subtraction and Shadow Detection "* IEEE Frame Rate Workshop 1999.
10. P. Rosin and T. Ellis *"Image Different Threshold Strategies and Shadow Detection'* 6th British Machine Vision Conf., Birmingham, pp. 347-356 1995
11. H. Bischof, H. Wildenauer and Ales Leonardis , *"Illumination Insensitive Eigenspaces"* ICCV, July 2001.
12. D Jacobs, P Bellhumeur, and R. Basri , *"Comparing images under variable lighting"* pages 610-617 CVPR 1998.