Technical Report TR-February-2009/02

Curtin University of Technology

Department of Computing

Multi Cue Performance Evaluation Metrics for Tracking in Video Sequences

GLADIS SUBHA JOHN

Prof. Dr. G.A.W. West

Senior Lecturer Dr. Mihai Lazarescu

14 February 2009

Multi Cue Performance Evaluation Metrics for Tracking in Video Sequences

Gladis Subha John, Geoff West, Mihai Lazarescu

Abstract

The key issue addressed by this paper is the necessity to devise performance evaluation measures for systems that integrate multiple cues for tracking in video sequences. We propose a generic evaluation approach that can be implemented in systems that perform higher-level people tracking by integrating multiple low-level features extracted from the video data. Two new measures: video sequence accuracy (VSA) and voting average measure (VAM), are introduced and explained by using the two fundamental image processing techniques: edge and optical flow detection. The effectiveness of the approach is demonstrated using a set of real video sequences with ground truth.

1. Introduction

This paper proposes methods to quantitatively assess the performance of a number of features pertaining to the detection, tracking and recognition of moving human beings e.g. pedestrians. The research is motivated by the need to revisit approaches to moving object analysis that consist of a hierarchy of stages with feature detection at the bottom and models of activity and human movement at the top. Further motivation is the emerging need to consider a wide number of different feature detectors running in parallel to get good segmentation by differentiating people from background. In the past, a number of methods have emerged that not only judge the quality of tracking systems but also compare the various approaches in order to measure improvements for existing approaches. These methods typically rely on one method e.g. background subtraction [30] or more recently the Scale Invariant Feature Transform (SIFT) [31]. Though the background subtraction method is very popular among the vision community it suffers from many flaws [4]. SIFT has recently been used by many because of its quite good performance for detecting different patterns over a range of angles and scales [2]. It can be argued that these and other methods are very much treated as "black box" processes i.e. they are used with little analysis of their performance in the context they are used in.

The assessment of various features is a much researched topic in recent years. Performance of edges [5], lines [25], texture [28, 10], corners [16, 19] and also methods like optical flow [1] have been evaluated by various authors. The key issue is that most of the evaluation studies are biased towards a single aspect or very much concentrated on single images [3, 12]. According to [24], technologies that involve multiple cues, which are limited only with regards to available resources, have advantages in computer vision. Methods have been proposed that integrate multiple cues [6, 13, 24]. Their reason for this is that multiple features can overcome limitations that each single feature has. Hence, there is much to be gained by evaluating features, finding those most suited, in combination, for the particular task and integrating them. The performance evaluation of such an integrated feature approach has rarely been looked into in the past, and the aim of this work is to introduce a new generic evaluation technique that dynamically determines the effectiveness of multiple low level features used in people tracking.

Overall, there are two problems with most of the evaluation methods proposed in the past: (1) they tend to be restricted to individual features (such as edge detectors) [26] and (2) they generally tend to provide an analysis that focuses on either the spatial or the temporal aspect of the results [27].

The aim our work has been to addresses these two problems in order to enable very accurate evaluation while allowing for the usage of feature sets. Specifically, in this report we describe an approach that enables the performance evaluation of multiple features (possibly used in combination) while providing an analysis that considers both the spatial and temporal aspects of the tracking task. We emphasise the evaluation of features in the context in which they are used prior to integrating them for a particular tracking task. The proposed methodology takes into account the pixel, frame and video sequence information and lead us to derive two generic measures: Video Sequence accuracy (VSA) and Voting Average Measure (VAM) to qualitatively evaluate the features for integrating in a tracking task. Though the method uses the two-class classification, it differs from other papers [8] in the context in which they are used. We demonstrate the efficiency of this methodology by evaluating four edge detection algorithms to track a person's head and feet using two optical flow methods using the edges inside the

bounding box as ground truth. The measures were tested on several video sequences for which provide a detailed set of results.

This work lead us to update the original measures to handle the issue of context in the evaluation process. By context, we mean that there could be changes in the environment, camera movement, changes in the number of people etc. A number of measures have been proposed in the past but they generally share the assumption that these conditions are reasonably stable e.g. slowly varying illumination. In this report we considers another two measures ---- Video Sequence Precision (VSP) to measure the precision, and Average Object Boundary Fragmentation (AOBF) used to quantify continuity in the edge-based contours that delimit the boundary of the object/s, and a modified version of Video Sequence Accuracy (VSA) . We present an analysis of the segmentation performance of several well known edge detectors on the PETS dataset for which we make available both the results and the pixel level ground truth data (the ground truth for the PETS dataset can be obtained for free from the IMPCA website at http://impca.cs.curtin.edu.au/downloads.php) and in addition to introduce the new measures for precision VSP and AOBF in an attempt to improve our precision as compared to the previous work.

2. Background

Performance evaluation is an important issue as evidenced by the regularly occurring IEEE Performance and Evaluation of Tracking and Surveillance (PETS) workshop. Evaluations in computer vision are carried out on real, synthetic or pseudo-synthetic sequences of varying length with or without ground truth. Some research has involved an evaluation framework comprising appropriate error metrics and video reference data sets, for the operational range of video surveillance systems [21]. In contrast with existing evaluation systems, which in general attempt to measure the overall tracking including blob detection, their segmentation separated motion detection and tracking. The objective of their research was to find the appropriate error metrics to evaluate the segmentation quality, the amount of spatial-errors as a function of foreground-background contrast, and to assess the effects of morphological processing on the detection results. The motion detection ground truth was generated using real image sequences with super-imposed computer generated humans, with each individual video sequence having a varying foreground-background contrast. Error metrics were created using hit rate, miss rate, false attempts and correspondence changes. Their results illustrated and quantified how points of discontinuity of the velocity vectors limit the operational range of tracking using a linear prediction model.

The performance of people tracking systems has been evaluated and several measures developed such as accuracy, practical measures and event sequence based error measures [17]. The accuracy measures included (1) Cardinality measure (2) Durational accuracy measures and (3) Positional accuracy measures. They propose practical measures: frame cardinality and event sequence based error measure to overcome the effort required in obtaining ground truth for ideal measures.

Objective metrics have been proposed to evaluate the performance of object detection methods by comparing the output of the video detector with the manually edited ground truth sequence sampled at one frame/s [29]. They detected and classified the errors as correct detections, detection failures, splits/merges or false alarms. A set of statistics such as the mean and standard deviation were compared for each type of error. They evaluated five algorithms namely basic background subtraction, W4, Single Gaussian Model, Mixture of Gaussians and the Lehigh Omnidirectional 3D Tracking System. Their proposed method provided a statistical characterisation by measuring the percentage of each type of error and enabled the user to select the best algorithm for a specific application.

It has been stated that the most common approach in evaluation is to vary the parameters of the input images or the algorithm and then construct receiver operating curves (ROC) [12]. It has been pointed out that many papers have presented an analysis that is specific to edge detection and the performance is given as a number, for example the percentage of edge points detected. However, there is little further analysis of the sensitivity of performance to relevant factors such as the context of the edge. The authors use the concept by psychophysicists which measures the effect on performance of variables in terms of the equivalent effect of a critical signal variable. They compared the performance of two line detection algorithms by detecting the presence or absence of a vertical edge in the middle of an image containing a grating mask and additive Gaussian noise. Their methodology could be applied to any detection problem. The main requirements of an effective performance analysis and the examination of methods for characterising video datasets have been described [9]. It was proposed to perform quantitative assessments over a wide range of conditions to satisfy the requirement of a real video surveillance task and the need to improve efficiency by using ground-truthed datasets.

Most of the abovementioned evaluation methods are biased towards a particular aspect, specifically either the evaluation of a feature or a method. Such methods could give good results on single sequences. A more generic

approach in the sense that it could combine multiple features for evaluation has been described [15]. The work involved the use of two comprehensive measures for text object detection and tracking systems for which the ground truth objects are bounded by simple geometric shapes. However, the approach had a very restricted application range with the focus being on detecting and tracking text in video which involves objects of consistent size and shape across a sequence. We adopt and improve this approach [15] by also concentrating on the details inside a bounding box, i.e. all the details inside a bounding box are marked ground truth and is the region of interest (ROI). Such an approach to our knowledge has never been adopted before in ground truthing of a real image.

3. Video Performance Analysis Methodology and VSA and VAM measures

We propose a methodology to evaluate quantitatively the performance of video sequences in tracking pedestrians using different features using a new Video Sequence Accuracy measure (VSA). The aim of the measure is to provide an effective way of evaluating the accuracy of a tracker by converting the spatio-temporal analysis to a single value that reflects the tracking performance over an entire video sequence. What sets our work apart from previous measures is the detailed level at which the tracking is analysed. Rather than relying on bounding boxes, centroids or area coverage, our measure attempts to summarize the tracking at the individual feature pixel level (such as the edge level) which is far more detailed and relevant than previous measures. For example, given the task of tracking a person, we attempt to provide a detailed low level spatio-temporal analysis of the tracking rather than simply checking whether the difference between bounding box or centroid in the test data are within an acceptable error when compared with the ground truth data. The measure can be easily applied to any features and methods in tracking. The methodology involves four major steps.

In the first step the ground truth for the video sequence to be evaluated is generated. The ground truth depends on the type of feature selected and it involves manual determination or expert supervision of a semi-automatic algorithm.

In the second step, the feature detection algorithms to be evaluated are applied to the original video sequence. The results are then converted to binary form (feature, no feature) e.g. edge pixel, not an edge pixel.

The third step of the approach involves the use of a two layered analysis of the results returned by the feature detection algorithms when compared with the ground truth. First, a rough comparison is carried out using ROIs. Since the evaluation is aimed at tracking moving people, a ROI bounded by a box that encompasses each individual person moving in the scene is selected in both the ground truth images and the result images. The coordinates of the bounding boxes are recorded for each frame for the ground truth in the video sequence and compared to determine if it matches the equivalent region in the result sequence for evaluation. When the ROI analysis has been concluded, a more detailed evaluation is carried out which involves a one to one comparison of pixels in the frames of the ground truth and the resulting binary images. The four possible outcomes when performing a one to one comparison of binary images are based on the familiar two class classification scheme:

- An edge pixel ground truth detected correctly as an edge pixel in the test image (True Positive - TP)
- A background pixel in the ground truth correctly detected as a background pixel in the test image (True Negative - TN).
- An edge pixel in the ground truth wrongly identified as a background pixel in the test image (False Negative - FN).
- A background pixel in the ground truth wrongly identified as an edge pixel in the ground truth (False Positive - FP).

These are used to measure the accuracy of each feature detected across the frames.

In the fourth step, the accuracy of the feature detection approach using the VSA is calculated using:

$$\mathrm{VSA} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i}$$

where n is the number of ground truth frames.

Once the VSA has been computed the tracking is evaluated by the Voting Average Method (VAM). A correctly identified person or part of the person being tracked is assigned a vote of 1 or 0 if not correctly identified. Parts of a person considered are the head, hands and feet. The VAM is calculated for each feature to be tracked in the sequence using:

$$\mathrm{VAM} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{F}_i$$

where n is the number of ground truth frames. The tracking system with the best values of VSA and VAM is considered to have the best performance for tracking. A specific and detailed example of how the approach works is presented in the next section.

3.1. VSA and VAM Measures Results and Discussion

All the results we present in this paper were obtained using an implementation of the evaluation framework that was coded using OpenCV and Visual Studio 2005 development environment on a Intel Core 2 Duo CPU E6850 @ 3.00 GHz, 2.99 GHz, 1.96 GB of RAM operating on Windows XP. The video sequences were taken from the recordings of the indoor security cameras in Department of Computing, Curtin University of Technology. For all video sequences processed, the aim was to investigate established edge detection and optical flow methods to track a person in an indoor scene. In all cases the scene contains several other persons and the complexity of the tracking task is compounded by the fact that part of scene background is similar to the person's clothing and occlusion occurs. The edges were considered in the context of the edges of the person to be tracked and the tracking was based on the ability of the optical flow method to track the head and feet using these edges. Four edge detection methods: Canny [7], Sobel [22], Roberts [20] and Prewitt [18] were evaluated in combination with the Horn-Schunk (HS) [11] and Lucas and Kanade (LK) [14] optical flow methods. Each optical flow algorithm only considered the edge pixels for analysis reducing the time for processing. First the ground truth was extracted. The perfect edge contour of the person in the video sequence was obtained by using the Canny edge detector as the starting point and then correcting it by creating the ROI using a bounding box and cleaning up the unwanted edges, separating the fused edges from the background and manually joining the gaps between the edges inside the bounding box. The Canny edge detector was used for this because it is well known to be good for noise suppression and for the accuracy of edge location.

3.1.1 Example Performance Evaluation of Edge and Optical Flow Features

An example of the image and its ground truth obtained this way is shown in Figure 1 and 2 respectively. Each video sequence has 100 frames and ground truth was taken for every fifth frame to reduce the manual editing workload.

Figure 1. Target walking away from the camera

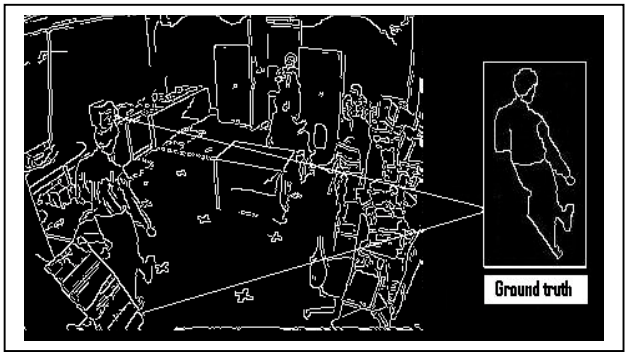

Figure 2. Ground Truth and the region of interest (ROI)

The edge detection using the four detectors was carried out on the test video sequence. Then a one-to-one mapping of the frames from each edge detector and ground truth frame was carried out after converting the images to their binary form i.e. '1' for an edge and '0' for the background. The one-to-one mapping was done only between the ROIs (Figure 2 and 3) in each frame. Then, a confusion matrix was created for each frame and the Video Sequence Accuracy (VSA) calculated.

Figure 3. Canny Edge Detector and the region of interest (ROI)

The next step in the process was to evaluate how the selected HS and LK optical flow methods track the head and feet using the edges. In order to do this, the video sequences containing only the edges were given as input to the two optical flow methods and a vote of 1 was recorded if it tracks the head correctly, and a vote of 1 recorded for each foot correctly tracked. If the tracking was incorrect a vote of 0 was recorded for each part. Voting was carried out for each frame in the edge sequence and then the VAM was calculated for the sequence. The results obtained are discussed in the following paragraph. The measures for each frame for the various edge detectors on the region of interest are tabulated in Table 1 for sequence 1. The graph of the accuracy of edge detected for the video sequence is shown as a graph in Figure 4.

Table 1. Accuracy of the edge detectors over the frames and VSA for Video1.

Frame No	ROI	Edge Detectors			
		Canny	Sobel	Roberts	Prewitt
18	23,85, 98,227	0.819	0.920	0.912	0.900
19	25,85, 99,226	0.830	0.920	0.912	0.902
20	28,83, 98,228	0.817	0.914	0.907	0.893
25	40,80, 107,226	0.822	0.908	0.902	0.890
30	52,75, 118,226	0.826	0.918	0.910	0.899
35	60,67, 119,208	0.790	0.885	0.866	0.886
40	66,63, 129,197	0.787	0.869	0.861	0.887
45	75,61, 135,199	0.811	0.863	0.891	0.879

50	84,55, 130,191	0.798	0.849	0.873	0.870
55	89,49, 136,169	0.795	0.841	0.873	0.871
60	94,46, 140,167	0.801	0.845	0.882	0.876
65	95,46, 141,166	0.810	0.866	0.888	0.884
70	100,40, 142,153	0.792	0.830	0.873	0.870
75	106,37, 149,142	0.797	0.825	0.868	0.866
80	114,36, 152,142	0.794	0.836	0.859	0.860
85	121,33, 154,130	0.787	0.842	0.843	0.846
90	12,632, 162,124	0.822	0.856	0.874	0.876
95	131,31, 166,124	0.818	0.851	0.875	0.881
100	133,31, 168,108	0.815	0.860	0.864	0.874
105	139,27, 171,111	0.795	0.828	0.843	0.849
110	145,26, 178,109	0.798	0.823	0.846	0.855
115	151,28, 187,111	0.809	0.826	0.866	0.862
VSA		0.806	0.862	**0.877**	0.876

The Roberts and the Prewitt edge detectors perform equally well for detecting the edges of the person to be tracked. It is worth noting that that even though the ground truth was taken using the Canny edge detector the results are not biased towards this detector that has the lowest VSA (0.806474) of all. This is mainly due to the false positives as the accuracy is measured for the edges to be tracked rather than considering all the detected edges. Moreover, the accuracy of all the edge detectors reduces as the distance between the camera and the person being tracked increases and vice versa. When leaving the room the accuracy fluctuates in each case.

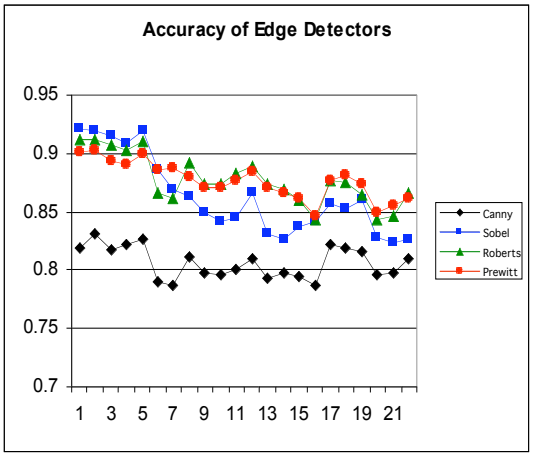

Figure 4. Accuracy of edge detection when the person moves away from the camera.

Figure 5. Target moving towards the camera

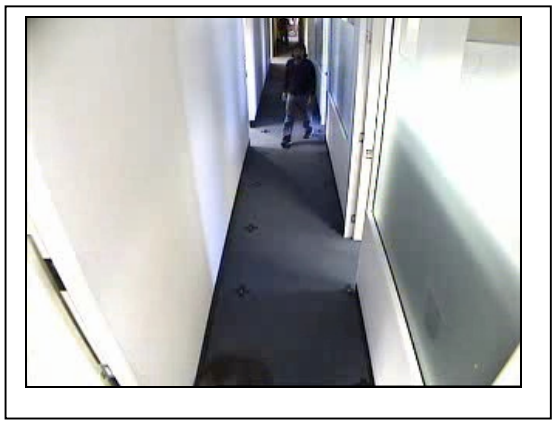

Figure 6. Target entering and maintaining a distance from the camera

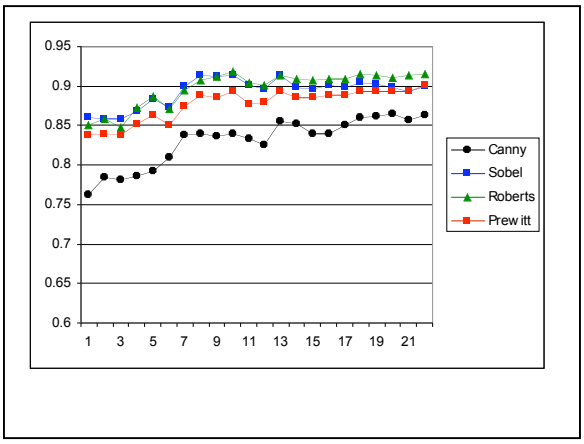

Figure 7. Accuracy of edge detection when the person moves towards the camera.

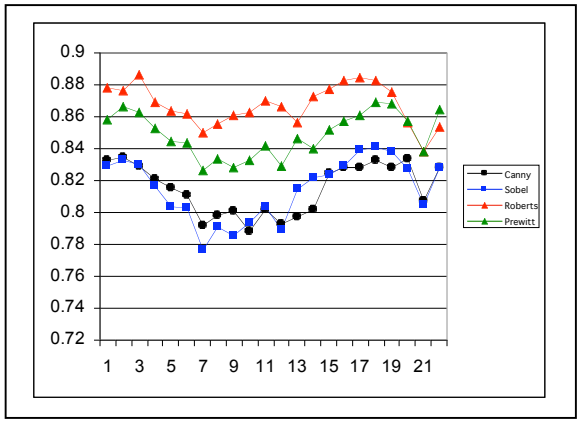

Figure 8. Accuracy of edge detection when the target maintains a distance from the camera.

Table 2. Voting Average Measure (VAM).

Video	True Positive Rate			
	Canny	Sobel	Roberts	Prewitt
1	0.863	0.909	0.909	0.94
2	0.863	0.505	0.386	0.636
3	0.845	0.845	0.904	0.88
Video	Precision			
	Canny	Sobel	Roberts	Prewitt
1	0.0777	0.081	0.075	0.079
2	0.0645	0.0727	0.0634	0.068
3	0.0427	0.0445	0.047	0.0445

Once the VSA values for the video sequences were calculated, they were evaluated for tracking the head and feet using the HS and LK method. The results for the VAM by voting for tracked head and feet for each frame are given in Table 2.

From the results shown in Table 2, it is noted that the VAM for Prewitt for tracking the head using the HS optical flow method is 1 and outperforms all the other detectors in tracking both the hands and feet. The second best method is Canny followed by either of the optical flow methods. The Prewitt efficiently detects the head in all the frames for which the ground truth was taken. Though the Roberts performs equally well in tracking the head using both HS and LK, it cannot be effectively used to track the feet. Sobel performs well to track the head using LK but suffers when using HS. Canny performs well with HS and exhibits average performance with LK. From the results depicted, the Prewitt edge detector with HS optical flow method is best to track the head and Canny with HS could be used to track the feet.

3.1.2 Additional Results

We have further tested our approach on another two (figure 5 and 6) video sequences and the summarized results are shown in Figure 7, Table 2 and Figure 8, Table 2 respectively. In the second video sequence, the target is walking towards the camera and exits the scene while in the third sequence, the target is entering the room, maintaining a distance from the camera before turning away and leaving the scene.

The results show that despite its simplicity, the Roberts edge detector is almost in all frames, the most accurate over the two sequences. Furthermore, the results confirm the earlier observation that the distance of the target object affects the accuracy of the tracking. In both cases, the VAM for Canny for tracking the feet using the HS and LK optical flow methods outperforms the other detectors. Similarly, the Roberts edge detector has the highest VAM values for tracking the head for both the HS and LK optical flow methods. It is worth to note that if the aim is to track the feet of a person, the HS optical flow when combined with the Canny edge detector produces consistently the best tracking results. Moreover, when tracking the head of a person, the HS optical flow method again generates the best outcome.

In addition to the VAM values, we also generated the precision and true positive rate for all the video sequences and the values are listed in Table 3. It can be noted that though the accuracy and true positive rate are reasonable the precision is very low. This is due to the fact that as only the edge pixels are considered as foreground pixel in the ground truth the number of false positives by running the detectors on the sequences will be very high and thus could impact the precision. As our main aim was to identify which edge pixels are correctly classified as foreground and background the precision doesn't have much impact and only the accuracy is considered for classification.

Table 3: True Positive Rate and Precision.

Optical Flow Method		Canny		Sobel		Roberts		Prewitt	
		Head	Feet	Head	Feet	Head	Feet	Head	Feet
1	HS	0.81	0.795	0.27	0.295	0.81	0.32	**1**	**0.795**
	LK	0.636	0.431	**0.81**	0.295	0.727	0.25	0.72	0.227
2	HS	0.909	**0.772**	0.545	0.363	**0.954**	0.363	0.863	0.59
	LK	0.772	**0.727**	0.863	0.545	**0.954**	0.5	0.818	0.59
3	HS	0.681	**0.772**	0.454	0.181	**0.909**	0.454	0.727	0.272
	LK	0.7722	**0.591**	0.772	0.272	**0.954**	0.636	0.909	0.363

Hence, the method requires higher edge detection rate of the person to be tracked to get good accuracy. This can be achieved by varying the threshold level of the detectors which is future research direction.

3.2. PETS2007 Dataset Results and Discussion

3.2.1 VSP and AOBF Measures

The first set of results lead to us to investigate how the changes in the context affect the measures described above. As a result, we propose another two measures (in addition to the VSA measure) to allow for a very detailed analysis of the performance of features while taking into the account a change in lighting conditions. Primarily we have attempted to address the key issue of the target boundary edge segmentation. Ideal edge segmentation would produce a set of connected edge points that match exactly those edge points in the target given by ground truth (GT). We argue that an effective way of determining the quality of the information provided by a feature can be summarized by a frame by frame precise analysis of the edges at the pixel level using three measures:

$$VSA = \frac{1}{N}\sum_{i=0}^{n}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$VSP = \frac{1}{N}\sum_{i=0}^{n}\frac{TP_i}{TP_i + FP_i}$$

$$AOBF = \frac{1}{N}\sum_{i=0}^{n}\frac{TP_i}{TP_i + FN_i}$$

where ***TP*** (true +ve) is a GT edge pixel of the object inside the bounding box correctly detected as an object pixel in the image, ***TN*** (true -ve) is a GT background pixel inside the bounding box correctly detected as a background pixel in the image, ***FN*** (false -ve) is a GT object pixel of the object inside the bounding box wrongly detected as a background pixel in the test frame, and ***FP*** (false +ve) is a GT background pixel inside the bounding box wrongly detected as an object pixel in the test frame.

The ***VSA*** and the ***VSP*** average the accuracy and precision computed per frame, over the sequence of ***N*** images. ***AOBF*** estimates the continuity of the object boundary by counting the number of correctly detected target boundary points as a proportion of the total number of target boundary points averaged over the ***N*** frames. For all three measures, the closer the values are to 1.0, the better is the performance.

3.2.2 Ground Truth Extraction

We used three videos were chosen from the PETS 2007 dataset S8 depicting an area near an airport terminal building. Three target scenarios are considered. They are:

Scenario	No of frames	GT
Target walking in the shaded region.	102	All frames
Target walking from the shaded region towards the sunny bright region.	125	Every fifth frame
Target walking from the sunny bright region towards the shaded region.	75	All frames ,BB for every fifth frame

Only one target was considered in a frame at a time even if there are multiple people in some frames. The region bounded within the box coordinates is considered as the region of interest (ROI). The scenarios are shown in Figure 9. The red line indicates the track for each target and the frames for which the ground truth are taken. The process in discussed in detail in the following sections.

Figure 9

Process I – Bounding box:

The bounding box values are given as spatial coordinates of the top left corner and bottom right corner. For example if the the bounding box values for frame 295 is 198,92,283,233, then the 198(x1) and 92(y1) indicates the corordinates of the top left corner abd 283 (x2) and 233(y2) indicate coordinates of the bottom right values respectively. The height and width of the bounding box can be easily obtained by:

$$\text{Height} = y2-y1 \text{ and Width} = x2-x1$$

The region inside the bounding box is selected as the region of interest for testing or evaluation purpose.

Process II – Edge Determination:

The next process is to determine the edges inside the bounding box that describes the target. In edge determination, Canny edge detector (Canny, 1986) was used in pre-processing as it is well known to be good for noise suppression and for the accuracy of edge location . The canny edge detector was processed to only produce the most significant edges of the target by eliminating short weak lists of edge pixels. This was obtained by recursively implementing the gaussian and its derivatives proposed by Deriche (1992). Gaussian filtering with a sigma value of 2.0 was used. The threshold for Canny was set to 750. Initially the ground truth reference edge maps were obtained using Canny and the edge strength of each edge point is computed in both directions perpendicular to the centre pixel by interpolation. Maximum suppression was applied to fix the position of the maximum edge strength and direction. The edge points were then linked together in lists of edge points and then sum the edge strengths for the edge points in one list and threshold. The reason for this is that in this way the long weak edge lists, the short strong edge list as well as the long strong edge lists will be detected and overcomes the issue of having a threshold that can fragment edges. The edges obtained this way are shown in figure 10.

Figure 10

The final step is to determine the edges of the target by cleaning up the unwanted edges, separating the fused edges from the background and manually joing the gaps between the edges inside the ROI. This was obtained by zooming the image in Microsoft Paint and care was taken to satisfy the human expertise of edge detection as discussed by (Salotti, Bellet, & Garby, 1996) and comparing pixel by pixel between the original frame and the processed frame. The end result is shown in Figure 11.

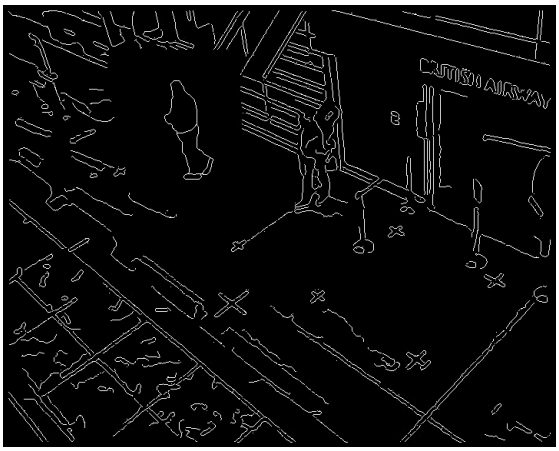

Figure 11

3.2.2 VSA, VSP and AOBF Results

Results for the three measures are shown in Table 4. Accuracy is high for all edge detectors over the three sequences showing good classification of edge and non-edge pixels. Precision is poor for all edge detectors because of many spurious edge pixels detected. Marr-Hildreth is the worst and along with Canny, is outperformed by Prewitt and Sobel.

Table 4. Results for the three measures.

Video	Video Sequence Accuracy (*VSA*)				
	Canny	Prewitt	Roberts	Sobel	Marr-Hildreth
1	0.912	0.938	0.927	0.938	0.867
2	0.926	0.935	0.925	0.934	0.852
3	0.957	0.960	0.950	0.960	0.930
Video	Video Sequence Precision (*VSP*)				
	Canny	Prewitt	Roberts	Sobel	Marr-Hildreth
1	0.328	0.444	0.292	0.443	0.179
2	0.429	0.538	0.363	0.529	0.210
3	0.540	0.658	0.415	0.657	0.460
Video	Average Object Boundary Fragmentation (AOBF)				
	Canny	Prewitt	Roberts	Sobel	Marr-Hildreth
1	0.787	0.560	0.461	0.553	0.411
2	0.803	0.439	0.319	0.435	0.494
3	0.833	0.570	0.278	0.559	0.460

Figures 12 and 13 show frame by frame accuracy and precision. The results are generally similar (except for Marr-Hildreth) and there is little fluctuation when the target is either in shadow or in sunlight. Better results occur when the target is in sunlight (increasing for video 2 and decreasing for video 3). In video 1, there is a sharp drop in precision near the beginning because the target is carrying a bag that is of the same color as the background at that region in the scene resulting in a reduction in detected target edge pixels. Canny produces less fragmentation of the edges than the other edge detectors meaning significant edge contours can be used for matching boundary parts. A reason for this is that it uses non-maximal suppression to keep valid edge pixels connected. Frame by frame fragmentation shown in Figure 14 shows that different illumination doesn't affect fragmentation significantly and Canny is consistently best.

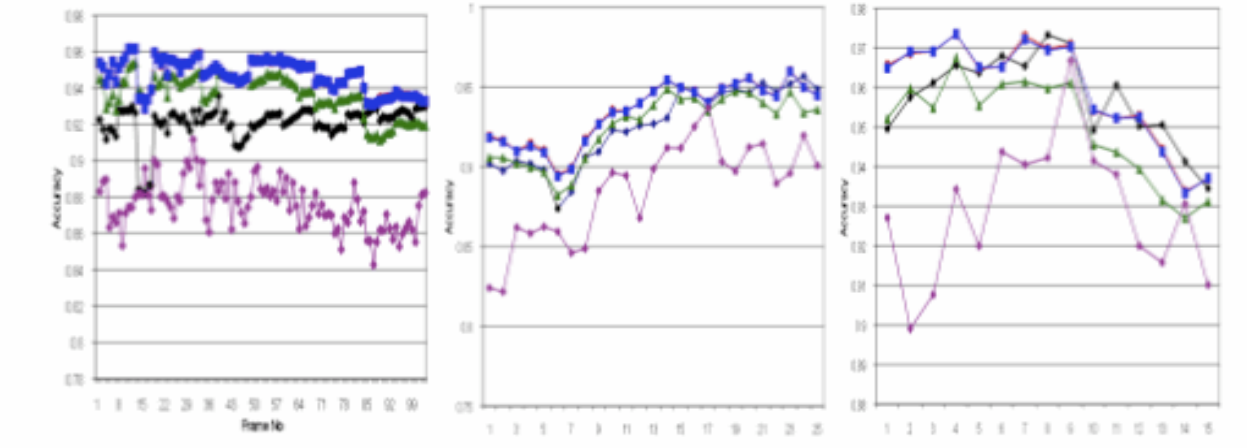

Figure 12. Target 1, 2 and 3 frame accuracy. Canny (black), Sobel (blue), Prewitt (red), Roberts (green), Mar-Hildreth (magenta).

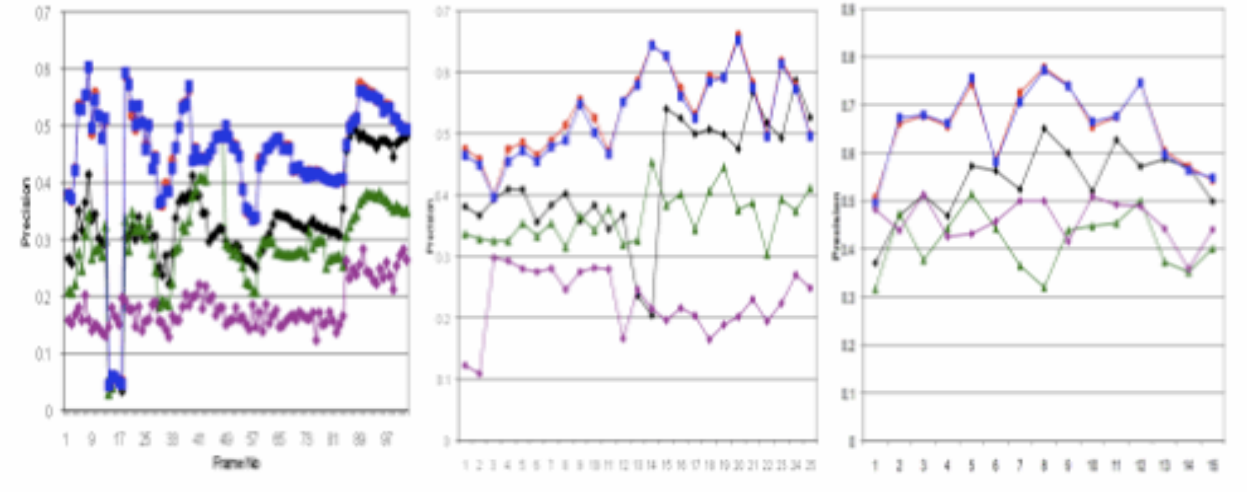

Figure 13. Target 1, 2 and 3 frame precision. Canny (black), Sobel (blue), Prewitt (red), Roberts (green), Mar-Hildreth (magenta).

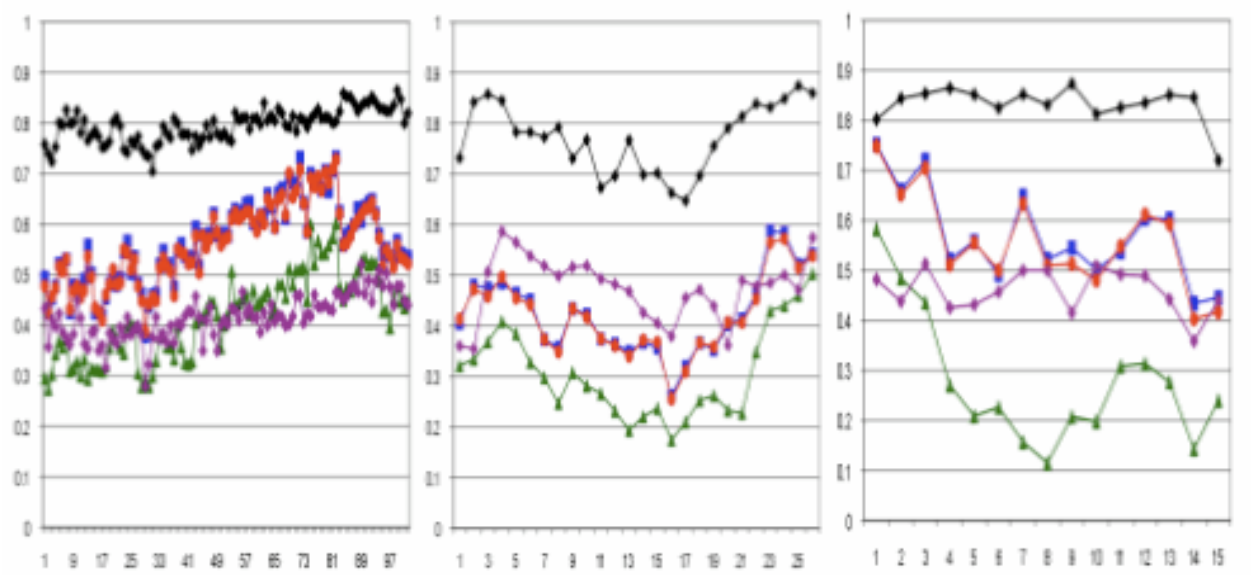

Figure 14. Target 1, 2 and 3 average object boundary fragmentation. Canny (black), Sobel (blue), Prewitt (red), Roberts (green), Mar-Hildreth (magenta).

4. Conclusion and Future Directions

The proposed methodology evaluates using pixel level, frame level and sequence level information and strongly demonstrates its application for choosing and evaluating various features and select the best features for the particular tracking task. The methodology is tested on two sets of video sequences. The first consists of three sequences of different orientations of the person to be tracked, under occlusion and also with little difference between the background and person's clothing. The second consists of three sequences from set S8, provided by PETS 2007.

The results for first set of sequences shows that Roberts and Prewitt edge detectors are the best followed by Sobel and Canny. From the test results for optical flow, the Prewitt edge detector with HS optical flow method is best to track the head and Canny with HS could be used to track the feet. For the algorithms used, what were considered the best parameters were used. Future work will look at the sensitivity to the parameters and how combinations of features, again goal directed can improve results.

The results for the PETS 2007 sequcnces show that the popular simple edge detectors considered are reasonable when used for segmenting targets (pedestrians) in surveillance video. There is some variation in results for the edge detectors when considering changes in illumination. The Marr-Hildreth doesn't perform well and this requires further investigation. Overall the Canny edge detector is best in the context of model-based methods because it suffers from less fragmentation of the significant edges producing richer features for matching. Future work will label the ground truth pixels with regions of the human body and investigate the performance of edge detection for the detection of these regions in the presence of occlusion and crowds of people. Other features (optical flow, texture, colour) will be investigated using similar techniques.This will enable the characterisation of these features and enable their integration into model-based muti-feature recognition and tracking algorithms.

6. References

[1]. Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1992). Performance of Optical Flow Techniques. *CVPR, 92*, 236 - 242.

[2]. Basharat, A., Zhai, Y., & Shah, M. (2008). Content based Video Matching Using Spatiotemporal Volumes. *Computer Vision and Image Understanding, 110*(3), 360-377.

[3]. Borra, S., & Sarkar, S. (1997). A Framework for Performance Characterization of Intermediate-Level Grouping Modules. *IEEE Trans. on PAMI, 19*(11), 1306 - 1312.

[4]. Bose, B. (2004). *Classifying Tracked Objects in Far-Field Video Surveillance*. Thesis. MIT. http://people.csail.mit.edu/cielbleu/pubs/Bose04MastersThesis.pdf.

[5]. Bowyer, K., Kranenburg, C., & Dougherty, S. (1999). Edge Detector Evaluation using Empirical ROC Curves. In *Computer Vision and Pattern Recognition*.

[6]. Brautigam, S., Eklund, J.-O., & Christensen, H. (1998). *A Model Based Approach for Integrating Multiple Cues*. Paper presented at the ECCV'98.

[7]. Canny, J. F. (1986). A Computational Approach to Edge Detection. *IEEE Trans. on PAMI, 8*(6), 679-698.

[8]. Dougherty, S., Bowyer, K. W., & Kranenburg, C. (1998). *ROC curve evaluation of edge detector performance.* Paper presented at the International Conference on Image Processing, Chicago, USA.

[9]. Ellis, T. (June, 2002). *Performance metrics and methods for tracking in surveillance.* Paper presented at the 3rd IEEE International Workshop on PETS, Copenhagen, Denmark.

[10]. Ghomi, F., Palmer, P. L., & Petrou, M. (April, 1996). Performance Evaluation of Texture Segmentation Algorithms based on Wavelets. In *Workshop on Performance Characterisation of Vision Algorithms*. Cambridge, England.

[11]. Horn, B. K. P., & Schunk, B. G. (1981). Determining optical flow. *Artif. Intell., 17*, 185-203.

[12]. Kanungo, T., Jaisimha, M. Y., Palmer, J., & Haralick, R. M. (1995). A Methodology for Quantitative Performance Evaluation of Detection Algorithms. *IEEE TRANSACTIONS ON IMAGE PROCESSING, 4*(12), 1667 - 1674.

[13]. Lu, S., Huang, G., Samaras, D., & Metaxas, D. (2002, 5-6 Dec). *Model-based Integration of Visual Cues for Hand Tracking.* Paper presented at the Workshop on Motion and Video Computing.

[14]. Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Int. Joint Conf. Artif. Intell.*, 674-679.

[15]. Manohar, V., Soundararajan, P., Boonstra, M., Raju, H., Goldgof, D. B., Kasturi, R., et al. (2006). Performance Evaluation of Text Detection and Tracking in Video. *DAS06,* 576-587.

[16]. Mokhtarian, F., & Mohanna, F. (2006). Performance evaluation of corner detectors using consistency and accuracy measures. *Computer Vision and Image Understanding, 102*(1), 81-94.

[17]. Pingali, S., & Segen, J. (1996). *Performance Evaluation of People Tracking Systems.* Paper presented at the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96), Florida, USA.

[18]. Prewitt, J. M. S. (Ed.). (1970). *Object enhancement and extraction*. New York: Academic Press.

[19]. Rajan, P. K., & Davidson, J. M. (1989). *Evaluation of corner detection algorithms.* Paper presented at the Twenty-First Southeastern Symposium on System Theory, FL, USA.

[20]. Roberts, L. (Ed.). (1965). *Machine Perception of 3-D Solids*: MIT Press.

[21]. Schlogl, T., Beleznai, C., Winter , M., & Bischof, H. (2004). *Performance Evaluation Metrics for Motion Detection and Tracking.* Paper presented at the 17th International Conference on Pattern Recognition (ICPR'04).

[22]. Sobel, I., & Feldman, G. (Eds.). (1973). *A 3x3 Isotropic Gradient Operator for Image Processing*: John Wiley and Sons.

[23]. Triesch, J., & Eckes, C. (1998). *Object recognition with multiple feature types.* Paper presented at the ICANN'98.

[24]. Wang, J. J., & Singh, S. (2003). Video analysis of human dynamics--a survey. *Real-Time Imaging, 9*(5), 321-346.

[25]. Wenyiun, L., & Dori, D. (1997). A Protocol for Performance Evaluation of Line Detection Algorithms. *Machine Vision and Applications, 9*, 240 - 250.

[26]. Smith, K., Gatica-Perez, D., Odobez, J.-M. & Ba, S., (2005), Evaluating Multi-Object Tracking, Proc. CVPR Workshop on Empirical Evaluation Methods in Computer Vision, San Diego.

[27]. Brown, L., Senior, A., Tian, Y.-L., Connell, J. & Hampapur, A., (2005), Performance Evaluation of Surveillance Systems Under Varying Conditions, Proc. PETS'95.

[28]. Chang K.,, Bowyer, K. & Sivagurunath, M., (1999), Evaluation of Texture Segmentation Algorithms, Proc. CVPR, Fort Collins, page 1294.

[29]. Nascimento, J. & Marques, J. (2005), Novel Metrics for Performance Evaluation of Object Detection Algorithms, Proc. 1st ISR Workshop on Systems, Decision and Control Robotic Monitoring and Surveillance, Lisbon.

[30]. Stauffer, C. & Grimson, W., (1999), Adaptive Background Mixture Models for Real-time Tracking, Proc. Computer Vision and Pattern Recognition, Colorado Springs, pp. 2246-2252.

[31]. Lowe, D. G., (2004), Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110.