SPATIAL-TEMPORAL REASONING BASED ON OBJECT MOTION

M. K. Teal & T.J. Ellis*.
Bournemouth University, Department of Electronics,
mteale@bournemouth.ac.uk.
*City University, Information Engineering Centre,
t.j.ellis@city.ac.uk.

Abstract

This paper describes the continuing development of a system for tracking multiple man made objects, (typically vehicles) moving in a natural open world scene, where the detected motion is used to construct a structural representation of the scene. The system assumes no *a priori* knowledge of any structure within the image, but begins building a map of the scene on a frame by frame basis. The map shows regions in the image where vehicles are likely to be detected and regions where they are likely to become occluded. Tracking is complicated by the fact that the vehicles to be tracked are expected to be a large distance from the camera and as such will only occupy a small number of image pixels. The system has been tested using an input sequence of vehicles moving in a complex outdoor scene, where the vehicles undergo both full and partial occlusion.

1 Introduction

There are many civilian and military applications where it is important to interpret structural features in a scene for identification and tracking of man made objects moving within that scene. The open world scene however can be a complex image to analyse, particularly due to illumination variations within the image and the changing pose of the object, each of which complicate the frame to frame matching of objects moving within the scene. Feature based geometric model matching, [2, 4, 10] has been shown to be very successful for identifying and tracking objects moving within a open world image, where the objects to be tracked occupy a significant proportion of the image; however they are less successful when the object to be tracked is further away from the camera and hence only occupying a small proportion of the image. In this case it has been found that the matching of crude object descriptors is more robust, [1, 3].

2 Overview

This system uses a static camera and frame differencing technique for detecting motion in an image which has a relatively static background. Objects with a measured temporal consistency are tracked across successive image frames. Regions in the scene are identified with particular types of dynamic events, such as regions

BMVC 1996 doi:10.5244/C.10.57

containing movement (e.g. roads), regions where motion occurs over relatively long scales (e.g. car parks) and regions where the object seems to disappear or partially disappear (occlusions).

An updating process is used to ensure that a reliable estimate of the background reference image is maintained by the system. Motion cues are matched against tracked objects from the previous frames using a simple model of temporal continuity and a spatial-temporal reasoning process is used to infer image structure. Because of the sensitivity of the motion estimator to changes in scene illumination and motion due to wind etc, a tile-based method is used to detect scene motion based on the estimations of statistical variations. The system is implemented in two stages, firstly stage 1 performs the detection, identification and tracking of objects moving in the scene and the second stage performs the spatial-temporal reasoning, which builds up an interpretation of structural features within the scene.

3 Detection, Identification and Tracking

The detection, identification and tracking process is comprised of two integrated algorithms, namely (i) the image acquisition, motion detection and reference generation and (ii) the target identification and tracking. The image acquisition, motion detection and reference generation algorithm inputs digitised images and applies a median filter to reduce noise caused during the digitisation process. Initially the first image from the input sequence is used to provide a set of reference grey level statistics (mean and standard deviation) and a reference edge image is generated by convolving the input image with a Marr-Hildreth edge operator [7] and detecting the zero crossings.

Motion cues are generated based on the results of grey level statistical differences between consecutive frames of image data and the reference grey level statistics. These motion cues form regions of interest (ROI) within the image and focus the attention of the target identification and tracking process. A set of object descriptors are generated for each ROI along with a measure of 'edginess' which gives an initial indication as to whether that ROI contains a possible target. The tracking applies a set of dynamic constraints on the motion of ROI's to help solve frame to frame correspondence and increase confidence that a tracked object is a target.

3.1 Image Acquisition, Motion Detection and Reference Generation

The image acquisition, motion detection and reference generation algorithm provides motion cues for objects moving in a scene. The algorithm is implemented in three stages. Firstly images are filtered using a median filter and statistical analysis is performed on four by four pixel regions (image tile) in the filtered 512 by 512 image and for each image tile the mean and standard deviation of the intensity are calculated. A standard t-test is used to identify significantly different regions between

the image statistics of current frame and the image statistics of the reference frame, hence determining regions that may contain motion (motion cues).

Motion detection using a frame differencing technique requires a suitable reference image (i.e. the background) and initially the first image in the sequence is used to initialise the background estimate. The apparent motion detected in the image is stored each frame, forming a history of the observed motion in the image sequence. In order to adapt to illumination variations, an updating strategy is employed to maintain the validity of the reference image. This is done via a statistical analysis of the motion cues, since it is observed that the cue detection rate tends to increase as the background estimate differs markedly from the true background. Figure 1 below shows a typical image sequence where the system is trying to analyse and classify the detected motion.

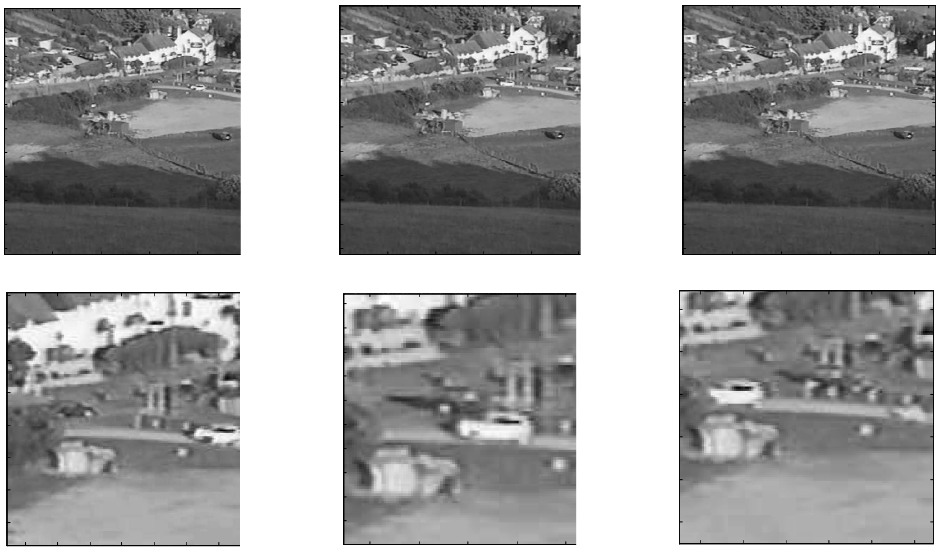

Figure 1: A three frame clip from an image sequence of over 90 frames. The top three images show a vehicle leaving the car park moving up a slight gradient in a left to right direction, at the same time a second vehicle is turning right and entering the car park. The car leaving the car park eventually occludes the vehicle that entered the car park. The lower three images show the corresponding enlarged portions of the original image where these vehicle are moving.

Every frame the update reference classification process is ‘triggered’ and the observed motion is analysed across a five frame window with the most recent five frames being used. The classification process compares the statistical results obtained for the current window with pre-determined limits on the size, type and number of objects moving within the scene. If the apparent motion is outside these limits then the

classification process updates the background with the current image. Figure 2 below shows the motion cues generated by the system for frames 2 to 69.

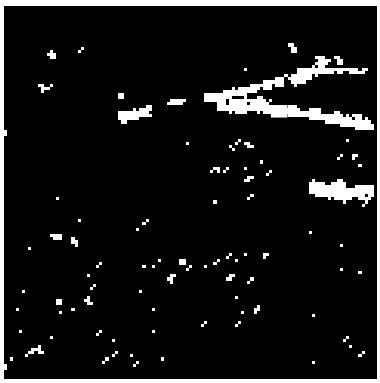

Figure 2: Motion cues generated
frames 2 to 69.

3.2 Target Identification and Tracking

The target identification and tracking algorithm identifies the regions of interest as either targets (man made vehicles) or objects (currently anything that is not considered to be a vehicle) and tracks these regions on a frame by frame basis. The tracking provides the spatial-temporal reasoning process with data on objects moving within the image, and is implemented in three parts. First the regions of interest found are segmented and a set of object descriptors calculated for each segmented region. Next edge analysis is performed on each segmented region and based on this analysis an initial identification of that region is made. Finally, dynamic motion constraints are applied to the segmented regions to resolve object correspondences, providing a further cue in the identification of a region.

Boundary regions are located using a two pass connected component labelling algorithm [6]. For each of these labelled regions the area and centroid co-ordinates are calculated using the zeroth and normalised first order moments and the min, max x,y co-ordinates are also determined. These calculations are all in tile co-ordinates which are easily translated back into a set of co-ordinates which define a rectangular bounding box in the original image (ROI). The frame number, number of objects for this frame, object descriptors and window co-ordinates are written into a object analysis table.

Initial target analysis is performed on each object in the table, in two stages, namely: edge extraction and initial target evaluation. The edge extraction is carried out using the Marr-Hildreth edge operator, which has the same standard deviation to that used in the generation of the reference edge image. The initial target analysis is attempting to identify internal geometric structure in a ROI that could be used to give an initial level of confidence that the region is a target. Man made objects could be assumed to consist of mainly straight line edges which occur infrequently in nature [8]. The edgels within the region of interest could therefore be used to provide an initial identification of that region (another cue in the identification process). Each

ROI is analysed and an ‘edginess’ measure calculated based on the ratio of edge pixels in the current image to edge pixels in the reference image. If the motion cue has been generated by an illumination change for example, then the edginess for that region is expected to be approximately unity as an edge detection operation is fairly robust to changes in illumination, in this case the ROI is initially labelled as an object. If the edge information in the ROI has change by a significant amount (a value of 10% has been found to provide good system performance) then this change is assumed to have been caused by a change in the structural features in the ROI (something has moved into that region) and is initially labelled as a target. Figure 3 shows the motion cues that where unmatched by the target identification and tracking system for frames 3 to 69.

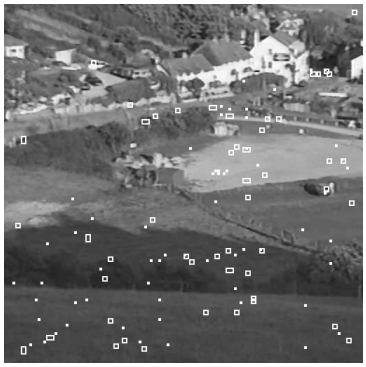

Figure 3: Un-matched motion cues frames 3 to 69.

However the edge information within these regions can be very sparse, consequently the edge analysis is only used to give an initial indication as to whether a region is a target or object. The results of the initial target analysis together with the object descriptors calculated for each region are written to an object description table. The generation of motion cues and the initial target analysis is repeated on a frame by frame basis, with target and object labels being generated each frame for every target or object found. This action forms a feature vector for each object in the image and across the image sequence an array of feature vectors are formed.

To solve the problem of object correspondence between frames, *a priori* dynamic motion constraints are applied. The maximum acceleration and orientation change of objects of interest with respect to the camera can be estimated *a priori*, based on the fact that objects to be tracked are distant from the camera and that these objects are rigid. The constraints are used to control a search algorithm that is attempting to minimise a Euclidean distance measure between object vectors in the current frame and those in the previous frame.

All target and object labels that satisfy the dynamic constraints are identified as targets and displayed. Target labels which have not satisfied the dynamic constraints are re-labelled as objects, and objects that have failed to meet the dynamic constraints are no longer processed. Figure 4 on the next page shows the matched and tracked targets extracted by system for frames 3 to 69.

Figure 4: Detected and tracked targets
frames 3 to 69.

4 Spatial-Temporal Reasoning

It has already been demonstrated [11, 12] that the motion of objects moving within an image can be used to construct some form of representation of that image. Here the spatial-temporal reasoning process is attempting to use the motion of objects moving within the scene, in this case vehicles, to form a structural interpretation of that scene. This interpretation takes the form of identifying areas within the image where vehicles can be expected to be observed moving and areas where vehicles could become occluded.

The interpretation process is split into two main tasks. The first task is to analyse the data supplied by the tracker, this data represents the trajectory of vehicles moving in the image together with a time index (frame number) and information about their size (area). The analysis groups the trajectory data into connected sets of segments which represent spatial areas in the image where vehicles have been observed moving (map segments). The second task takes the map segments from the spatial analysis process and applies a spatial reasoning process to the possible spatial and temporal relationships between these map segments.

4.1 Spatial Analysis

The spatial analysis takes the target tracking data output from the tracker and constructs sets of linear map segments based on the target trajectories. The linear segments are mapped into the image in tile co-ordinates and linked using an 8-neighbourhood connectivity algorithm. Each connection made between individual linear segments has an edginess factor calculated for that segment. A time index is included, based on the frame number, and an observation factor is calculated from the number of instances, that targets have been observed in that segment. The structure of a map segment is shown on the next page in figure 5.

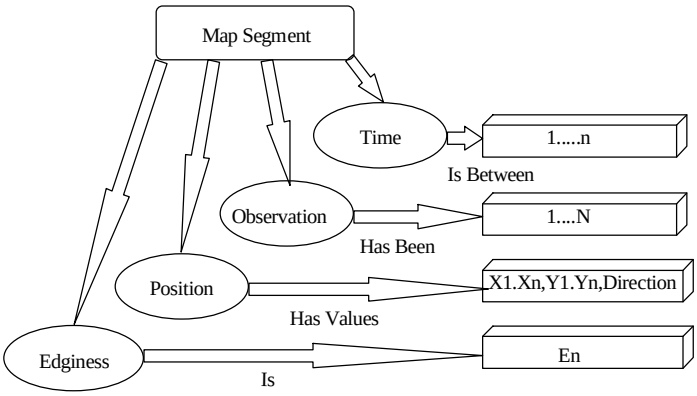

Figure 5: Map Segment.

4.2 Spatial Reasoning

The spatial analysis process generates map segments in tile co-ordinates that describe areas within the image where we have inferred motion; these map segments form the building blocks for the spatial reasoning process. No assumptions are made about any *a priori* structure within the scene, so initially all regions in the map are labelled as being 'unknown', thus the system effectively starts with an empty map. The reasoning process takes the map segments and using a set of rules, infers the most likely interpretation for a region. This inference mechanism is structured using a semantic network, which is shown in a simplified form in figure 6. The network consists of four arcs, namely 'part of', 'add to', 'next to' and 'between' and five object nodes 'road segment', 'ground segment', 'road', 'ground' and 'static'.

'Part of' takes a map segment and checks to see if it is part of a road segment or part of a ground segment, invoking a set of spatial and structural operators to accomplish this task. If the identified segment is a repeat of a map region, then 'add to' adds it to that map region, if the identified segment has not been observed before then 'add to' generates a new map region for that segment. The labelled regions each have a confidence factor associated with the label, and this factor is increased each time motion is observed within that region. The entire map is then scanned and regions that have been identified as either road or ground are checked to see if any of these regions are 'next to' one another. This operation enables regions of the map that have been identified as areas where vehicles can be expected to be observed moving, but are not thought to be roads to be linked to a road (a dirt track may join a road at a junction for example).

The 'between' operation applies a set of geometric rules that uses the premise that roads or ground regions are associated with motion and can be linked using straight lines (roads are considered to be straight) within a search space. If links are established between identified regions, those links are labelled as static, i.e. that area of the image could contain an object that may occlude vehicles moving in the image. However if motion is observed in any spatial links established between identified regions, the region is re-labelled as either road or ground.

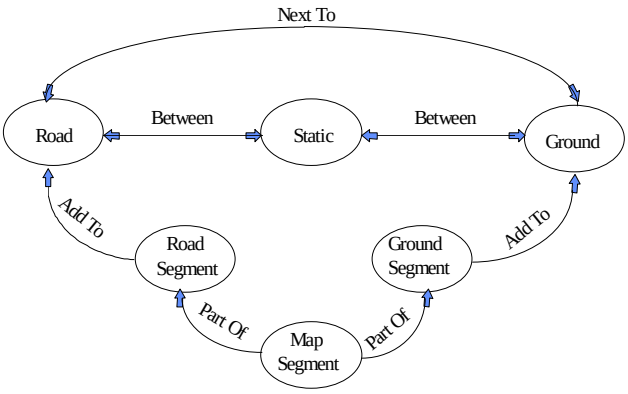

Figure 6: Simplified semantic network for spatial reasoning of map segments.

5. Discussion

A static cam-corder was set up and an open world image sequence filmed showing vehicles and people moving in that scene. From this sequence a 90 frame clip was digitised to disk at a rate of approximately two frames a second. The detection, identification and tracking process showed that the statistical analysis removed most of the false motion cues generated by using a frame differencing technique to perform the motion detection, and that sufficient resolution still remained to detect and track the vehicles moving in the image despite the fact that some of these vehicles occupy an area of less than 100 pixels. Techniques that use frame differencing to determine motion between consecutive frames of image data require a reference image that must first be acquired and possibly updated in some form, [9]. In this research, the method of reference update is determined by the number of motion cues detected in the image across a sliding window. Figure 7 shows the rate of cue detection over the entire frame sequence, which increases as the image sequence is processed.

The increase in cues is due to false cues being generated by changing illumination conditions, (the sequence was filmed early evening with the sun setting behind the camera). At frame 75 the classification process updates the reference image data generating new statistical and edge data by simply taking the current image frame. After the new reference data was generated, the drop in the number of motion cues perceived in the image was reduced from 96 down to 3. Updating the reference had no effect on the constructed map, and a van observed at frame 84 leaving the car park and turning right onto the main road where it would become occluded, was identified and tracked by the system.

The initial identification process together with constraints placed on the motion of the objects moving in the image removed most of the false motion cues whilst still tracking actual targets. Figure 4 shows the identified and tracked motion cues (vehicles) superimposed on the original image. The tracking system cannot at present resolve the problem of multiple target tracking with tracked objects either partially or fully occluding each other, however, Toad et al, [13] has shown that reasoning strategies can be used to overcome problems with this type of occlusion.

The map segments generated by the spatial analysis process define the regions within the image where vehicles have been detected.

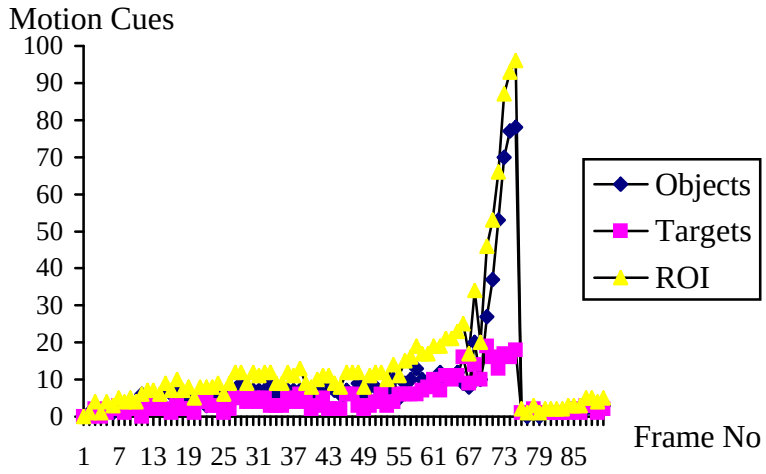

Figure 7: Plot showing the total number of motion cues found per frame, classified into either targets (vehicles) or objects.

The spatial reasoning process has grouped these segments to define areas in the image where targets are expected to be observed moving. The spatial extrapolation using the 'between' premise identified two areas where targets could undergo occlusion. The map is constructed on a frame by frame basis and figures 8a, 8b and 8c below show a 2-D representation of the map in tile co-ordinates, as it was constructed (the system is learning about structural features in the scene) for frames 15, 35, 55. White signifies areas where vehicle motion has been detected and in future where the system should expect to observe further target motion.

Figure 8(a): Scene Map after 15 frames.

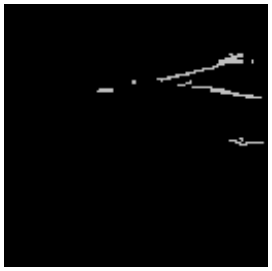

Figure 8(b): Scene Map after 35 frames.

Figure 8(c): Scene Map after 55 frames.

The map effectively represents contextual-information about the scene layout, which could now be used to improve the target tracking [14], by focusing the processing of the system to those areas expected to contain target motion.

6. Conclusion

The system demonstrated that it is capable of extracting and tracking man made objects (vehicles) moving in an open world image sequence, when the tracked vehicles are a large distance from the camera. The extracted motion data was sufficient to construct a map that represents areas of the image where vehicles can be expected to be observed moving and the use of simple spatial extrapolation rules, areas in the image were determined where vehicles could become occluded.

Future work is aimed at using this constructed map to improve the tracking of targets by focusing the attention of the image processing algorithms to those areas expected to contain motion and to develop a strategy to alternate between areas of high activity and areas of low or no activity. A strategy is also under development for the predication of target occlusion so that when vehicles become occluded in the image, but are still in the field of view of the camera, they can still be tracked.

References

1. Teal MK, Ellis TJ, “Target tracking in open world scenes using motion cues and target dynamics”, IEE 5th International Conference on Image Processing and its Applications, 1995, pp 276-280.
2. Tan TN, Sullivan GD and Baker KD, "Fast vehicle localisation and recognition without line extraction and matching", Proc. BMVC 1994, **1**, pp 85-94.
3. Rosin PL, Ellis TJ, "Detecting and Classifying Intruders in image sequences", Proc. BMVC 1991, pp 293-300.
4. Worrall AD, Sullivan GD, Baker KD, "Advances in model-based traffic vision", Proc. BMVC 1993, **2**, pp 559-568.
5. Jain R, "Difference and accumulative difference pictures in dynamic scene analysis", Image and vision computing 1984, **2**, pp 99-108.
6. Image processing, Analysis and machine vision. Sonka, Hlavac, Boyle, 1993.
7. Marr D & Hildreth E Proc R. Soc Lond. B 207 ,1980, pp 187-217.
8. Radford CJ, "Vehicle detection in open-world scenes using a Hough Transform Technique", IEE 3rd International Conference On Image Processing & Applications 1990.
9. Rosin PL, Ellis TJ, “Image difference threshold strategies and shadow detection”, Proc BMVC 1995, **1** pp 347-356.
10. Koller D, Daniilidis & Nagel H-H, “Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes”, International Journal of Computer Vision, **10:3**, 1993, pp 257-281.
11. Li-Qun X, Hogg D, “Building a Model of a Road Junction Using Moving Vehicle Information”, Proc British Machine Vision Conference 1992, pp 443-452.
12. Johnson N, Hogg D, ”Learning the Distribution of Object Trajectories for Event Recognition”, Proc BMVC 1995, **2**, pp 583-592.
13. Toal A. F, Buxton H, “Spatio-temporal Reasoning within a Traffic Surveillance System”, 2nd European Conference on Computer Vision 1992, pp 885-892.
14. Gong S, Buxton H, “From contextual knowledge to computational constraints”, Proc BMVC 1993, **1** pp 229-238.