AUTOMATIC HUMAN BEHAVIOUR RECOGNITION AND EXPLANATION FOR CCTV VIDEO SURVEILLANCE

September 12, 2006

Abstract

This paper is concerned with producing high-level text reports and explanations of human activity in video from a single, static camera. The motivation is to enable surveillance analysts to maintain situational awareness despite the presence of large volumes of data. The scenario we focus on is urban surveillance where the imaged person is medium/low resolution. The final output is text descriptions which not only describe, in human-readable terms, *what* is happening but also *explain* the interactions which take place. The input to the reasoning process is the information obtained from lower-level algorithms which provide an abstraction from the image data to qualitative (i.e. human-readable) descriptions of observed human activity. Causal explanations of global scene activity, particularly where interesting events have occurred, is achieved using an extensible, rule-based method. The complete system represents a general technique for video understanding which requires a guided training phase by an experienced analyst.

1 Introduction

A system which could automatically report on human activity in video would be extremely useful to surveillance officers who can be overwhelmed with increasingly large volumes of data. In both the civilian and military domains, the maintainence of situational awareness is critically important. This is due to the fact that, when an analyst focusses attention on a specific object of interest, potentially he/she is unaware of other interesting, suspicious or dangerous activity in the same scene. This problem is exacerbated when multiple screens must be monitored. Moreover, a system which could subsequently explain this activity would be a significant development in the technical area of video-based human behaviour understanding.

Computer Vision researchers, however, generally have focussed on developing lower-level techniques for analysing image sequences, such as feature-tracking

and face/skin detection e.g. [23, 25]. Whereas the Artificial Intelligence community has contributed to the problem of expert knowledge representation and human-like reasoning processes e.g. [6, 12]. However, it has been recognised that there is a distinct lack of attempts to develop a system for visual scene understanding which combines the necessary aspects of both disciplines for intelligent visual surveillance [15]. The work of this paper addresses this need without excluding the "man-in-the-loop". In fact, we utilise expert prior knowledge to ensure that the output descriptions on activity are accurate and to reject spurious targets. To that end manual input is used to define the rules for the higher-level processes governing human activity and to provide the training data labels. (In a simple urban surveillance scenario these qualitative descriptions might include, for example, *nearside-pavement*, *on the road*, *far-side pavement* for position, *left-to-right*, *away*, *towards* etc. for direction.)

This work goes beyond simply reporting on individual activity, albeit at a human-readable level: in this paper we also present a prototype system for *reasoning* about human activity in video. The system is split into two main parts: (i) low-level activity recognition for a single person in video which is described in section 4; (ii) higher-level reasoning about events using this information which is described in section 5. We demonstrate the efficacy of our system by presenting results from the urban surveillance domain (although the techniques are equally applicable to further applications e.g. sports footage as we show in other published work [16, 17]).

The remainder of this paper is structured as follows. We begin with a review of the relevant prior art, then turn to a more detailed description of each of the stages of the method. We describe the data and the role of the "expert" operator in section 3. A technique to estimate where a person is looking is described in section 4.1. Single person action recognition is described in section 4.2. Sequences of action comprise the overall behaviour of an individual, and we represent these sequences using stochastic models in section 4.3. Together this qualitative information is defined as the information available to the "sensors" of a human agent (mainly, in this paper, a *pedestrian* agent). Rule-based reasoning, using rules derived from expert knowledge about the domain, is introduced in section 5 to generate human-readable text explanations of the observed activity. The final result, in section 5, is therefore not only a high-level description of all scene activity but a causal explanation of interesting events. We conclude in section 6 and discuss avenues for future research in section 7.

1.1 Related work

There has been much reported in the recent literature about methods for training recognition systems using large training data sets (e.g. [23]). Recently Zhong *et al.* [26] demonstrated detecting unusual activity by classifying motion and colour histograms into prototypes and using the distance from the clusters as a measure of novelty. Also Zelnik-Manor and Irani [25] used a distance metric to identify examples of actions in video. Boiman and Irani [3] address the problem of detecting "irregularities" in video, where "irregular" is defined solely by the

Figure 1: Gaze-direction is an important clue to intention.

context in which the video takes place. Xiang and Gong have addressed an important issue: how to effectively recognise action in a surveillance context when there is a sparsity of example data [24] and what rôle the high-level labelling of trajectories plays in this situation.

Making sense of a scene can be thought of as, "Assessing its potential for action, whether instigated by the agent or set in motion by forces already present in the world" [4]. In other words, a causal interpretation is most easily and most commonly judged by the motion effects that take place. Michotte, with Heider & Simmel demostrated that it is the kinematics of objects that produce the perception of causality, not appearance [19]. There is, nonetheless, a history in scene understanding research of analysing static scenes. In the work of Brand and Cooper [4]. One major shortfall in the reported work on reasoning, from an Artificial Intelligence perspective, is the lack of robust computer vision methods for obtaining low-level information about complex visual scenes and the agents within them [15]. The work of Brand *et al.* relied on the extraction of very simple visual features from static images of blocks against a white background. Siskind demonstrated reasoning about the dynamic interactions between tracked blobs (hands, blocks) in simple video sequences [22]. Our work addresses this gap by applying established techniques to generate probabilistic estimates over qualitative descriptions of human activity in video [16, 17].

"Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" is an agent, according to Russell and Norvig [18]. An agent is, therefore, analogous to a software function. When human agents are combined, complex behaviour emerges which can model real-world behaviour as demonsrated by Andrade and Fisher for simulated crowd scenes [1]. There are many types of agent defined in the AI literature. The Belief-Desire-Intention agent is believed to model decision-making process humans use in every day life [10]. Related to agents, and of considerable relevance to the work of this paper, is the work of Dee and Hogg [8] in which a particular model of human behaviour is verified by comparing how "interesting" the model indicates the observed behaviour is to how worthy of further investigation a human believes the behaviour to be. Dee and Hogg's work focusses on inferring what an agent can sense through line-of-sight projection of rays and the subsequent use of a predefined model of goal-directed behaviour to predict how the agent is expected to behave. Not all of the information required

for reasoning is automatically extracted from the images (which is an area we explicitly address in this work).

On rule-based reasoning, Siler notes that rules have, "... shown the greatest flexibility and similarity to human thought processes ..." [21]. These rules can quickly be identified and written down by an expert. A significant positive aspect of rule-based reasoning is that it is easy to update the system's knowledge by adding new rules without changing the reasoning engine [15]. It is also easy to transfer between applications by specifying a new set of rules.

2 Contributions

In relation to the prior work in this area, the contributions of this paper are:

- Our method requires much less training data than the common human activity recognition techniques found in the literature [5] which are based on statistics (i.e. hours vs. days),
- We explicitly use the prior knowledge of an analyst familiar with the scenario which is not only technically advantageous (i.e. provides more accurate results), but is strongly aligned with the needs of surveillance professionals,
- Our system achieves "human-like" reasoning about causal relations between imaged people direct from an input video by way of complex, dynamic visual features, which has not been demonstrated until now.

3 The role of the operator

As opposed to automatically tracking and identifying every moving object in the scene (which may, in the future, be fully implemented), we utilise the knowledge of an analyst to select objects of interest during operation of the system. In practice this enables the analyst to initiate an automatic information-gathering process for the objects of interest while, at the same time, remaining visually aware of other activity in the scene. The expert analyst is considered to be very familiar with the scene and the type of language which should be used to report on the scenario in his/her domain.

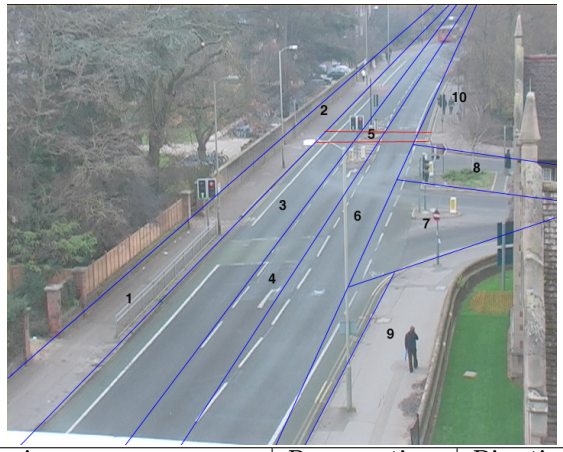
The use of an expert has two important results. The first is that nuisance and false alarms are considerably reduced since the identification of targets is achieved using manual intervention, considerably which eradicates clutter due to false alarms. The second is that the analyst can assist with the gathering of training data. An example of a scene marked-up by an analyst is shown in figure 2. The analyst will further label examples of specific activity which have been selected for the training database. This entire process typically takes less than 1 hour for a new scene.

In this paper we consider 2 different scenes for analysis. In figure 1 we show the relative quantities of training data required for each of these scenes. Note

Sequence	Total (frames)	Training (frames)	Testing (frames)
Urban street	5455	665	2361
Junction surveillance	76040	4491	18445

Table 1: The data volume for each of the videos used in the analysis of our technique described in this paper.

that a short training phase is required for every new scene. Training involves the analyst selecting an example of normal activity and providing a suitable label, for example "Walking, North Pavement". The data and associated labels from this guided training phase are used to automatically generate descriptions later.



Regions	Person actions	Directions
Northbound Lane (3)	Walking	North
Right Turn Lane (4)	Running	South
Southbound Lane (6)	Stopped	East
Parks Road Westbound (7)		West
Parks Road Eastbound (8)		Stopped
South-East pavement (9)		
South-West pavement (1)		
North-East pavement (10)		
North-West pavement (2)		
North pedestrian crossing (5)		

Figure 2: The scene is divided into regions and labelled by an expert analyst. The labelled regions, activities and directions for this scene are detailed in this table.

4 Single-person activity estimation direct from video

As we have stated, the overall goal is that we may be able to automatically explain human activity in video. The information we require to achieve this goal becomes apparent when we consider what a human might need to know to *reason* about human activity e.g. What are the people doing? What can they sense?

In contrast to the previous work in this area, the video processing techniques we have developed and proven in earlier published work [16, 17] answer these questions automatically while retaining some knowledge of uncertainty using probability theory [11]. The following information is extracted directly from the video:

1. Gaze-direction: what lies in a person's visual focus of attention, in the image.
2. Spatio-temporal action: e.g. walking or running.
3. Behaviour: a sequence of spatio-temporal actions e.g. crossing the road.

These activity descriptions taken on their own comprise a *report* of the video. We then use this report to *explain* observed interactions using a rule-based reasoning approach. In this section we describe the probabilistic activity estmation: gaze-direction estimation (section 4.1); spatio-temporal action recognition (section 4.2); and finally behaviour recognition (section 4.3).

4.1 Gaze direction estimation

The first lower-level component of our system estimates where a person is looking in images where the head is typically in the range 20 to 40 pixels high, which is representative of most CCTV-style footage [17]. We use a image features derived from detecting skin pixels in static images to estimate the orientation of the head, which is discretised into 8 different orientations, relative to the camera. A fast sampling method returns a probability distribution over previously-seen head-poses. The overall body pose relative to the camera frame is approximated using the velocity of the body, obtained via automatically-initiated colour-based tracking in the image sequence [7]. By combining direction and head-pose information gaze is determined more robustly than using each feature alone. We show an example of this process applied to surveillance footage in figure 1.

Our results from across a range of test sequences indicate we can achieve gaze-direction estimation with a median error of 5.5^o using this method when applied to faces at the resolution of those shown in figure 1. The main source of error is due to the resolution of the head-pose estimate. (However, the angle of projection of the scene onto the image can cause some errors also, as we discuss in [17].)

Figure 3: Action recognition is achieved by exploiting the dominant patterns of human motion (i.e. arms/legs) within a target-centred window. Here, we show input frames (top row) with corresponding database matches (bottom row).

4.2 Spatio-temporal actions

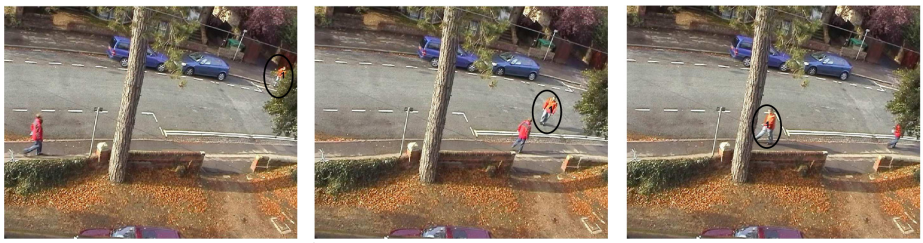
In addition to gaze-direction we also require to extract basic information such as position, velocity and activity-type e.g. walking vs. running vs. standing etc. To that end we employ a technique for sampling from hand-labelled exemplar databases [20]. This sampling method returns a probability distribution over a set of predefined examples, where the qualitative labels of place, direction and action-type have been identified by an expert user as described in section 3.

This method holds three significant advantages: (i) high-level descriptions can be incorporated by a qualified expert; (ii) by sampling from the data, far less training data is required than is the case for standard, statistic-based learning techniques (see, for example, [5]); (iii) probabilistic distributions prevent us committing to one interpretation of activity too early.

Position and velocity example data is derived directly from the centroid of the object as estimated using a colour-based tracker [7]. This tracking solution provides reliable centroid data during the life-time of an object within the video since weather and lighting variability is not large while the target is under observation. Although some consideration needs to be given to this issue in further studies. Action-type is encoded using a descriptor based on 2-D image velocity [2]. The feature we use to discriminate between human actions is an extension to the descriptor of Efros *et al.* [9] and is based on extracting the gross properties of motion from the optic flow field. An example of matching actions using this method is shown in figure 3.

The position, velocity and action-type image feature databases are maintained independently. This enables more efficient use of each feature, reducing the volume of training data required. Fusion of each of these independent features allows us to compute probability distributions over *spatio-temporal* actions from the independent distributions over the feature databases. As a simple example, the spatio-temporal action "walking on pavement" is created from the place "pavement", and the action "walking" (with a certain direction).

It is important to note that this fusion step retains the full probabilistic



Frame	Activity	Likelihood
1 - 70	Walking on far-side pavement	0.86
71 - 225	Walking on road	0.94
226 - 450	Walking on near-side pavement	0.94

Figure 4: An accurate commentary is obtained for this urban street scene where the person circled is under observation. Note that partial occlusion does not have a detrimental effect on performance.

Frames	Estimated activity
5330-5470	road, running

Figure 5: A second example from a more challenging surveillance scene.

information at our disposal. That is, the *probability distributions* over the position, velocity and action features are fused using a small Bayesian network. (More details of this step can be found in [16].)

By taking the most probable estimate[1] of all possible spatio-temporal actions at each time step, a commentary on activity is generated. An example of this commentary for surveillance video is shown in figure 4. On the basis of frequency of occurrence, the prior belief for each spatio-temporal action can be derived directly from the training datasets. They can also be easily hand-tuned. In the commentary example of figure 5, the priors are critical to the probability computation and subsequent choice of the correct spatio-temporal action. Running is not represented as often in the example database. Therefore if the priors for each simple-action are computed on the basis of frequency then the most likely spatio-temporal action for this sequence is *road, walking*. If however, the priors are uniform the most likely result is as shown, which is correct.

Detection rate statistics for the two sequences under analysis in this paper are shown in the table of figure 2.

[1] We use the Maximum Likelihood (ML) estimate from the probability distributions.

Sequence	% true detection
Urban street	96.7
Junction surveillance	74.0

Table 2: The true detection rates for the three video sequences used in this chapter. We compare the automatic descriptions of action to hand-labelled ground truth.

4.3 Behaviour as a sequence of spatio-temporal actions

Having generated probability distributions over actions, we subsequently use Hidden Markov Models (HMMs) to encode known rules about behaviour. For more detailed information about HMMs see e.g. [14]. In brief, HMMs are fully defined by a set of state transitions, initial probabilities and output probabilities from each state. We are particularly interested in this type of compact model for two reasons: (i) the transition matrix can efficiently encode known "rules"; (ii) inference using an HMM retains uncertainty which is crucial to human reasoning, especially when displaying information to an operator.

The spatio-temporal action is an abstraction from the image data (i.e. pixels) to a higher-level description of activity in the scene (i.e text). Taken on its own, it provides a readable commentary on activity. When thus abstracted to a human-level description it is no longer dependent on the particular camera viewpoint from which it was initially generated. This is significant as it enables us to map the high-level knowledge of the user to rules which facilitates further video analysis. This is done by encoding the probability of transitions between spatio-temporal actions. For example, the behaviour "crossing road" is composed of a sequence of spatio-temporal actions "walking on far-side pavement", "walking on road", "walking on near-side pavement", in that order, as shown in figure 4. The HMM allows this to be very efficiently encoded in the transition matrix of the model.

Once more, the operator is involved to provide an accurate example of each behaviour to be modelled. This example is then automatically "parsed" into a sequence of actions and the state transitions for that model are thus identified.

For the scene in figure 6 we encoded 3 such behaviour models very efficiently by defining the transition and initial-state probabilities for each model. These models correspond to the behaviour "crossing road", "walking along pavement" and "turning into drive".

On-line estimation of which model best explains the observed action-sequence enables us to estimate higher-level behaviour. A Likelihood Ratio is used for model-selection to alleviate the problem of higher-order models naturally providing a better explanation of the data.

Note that, even if the global behaviour is not recognised, a sensible description of activity can still be achieved from the action-recognition stage of our system. Also, since these behaviour models are general to the scene, they can discriminate between the same type of behaviour performed in different ways without the need for separate models (as a learning technique trained directly from the image data would require). An example of this feature in operation is

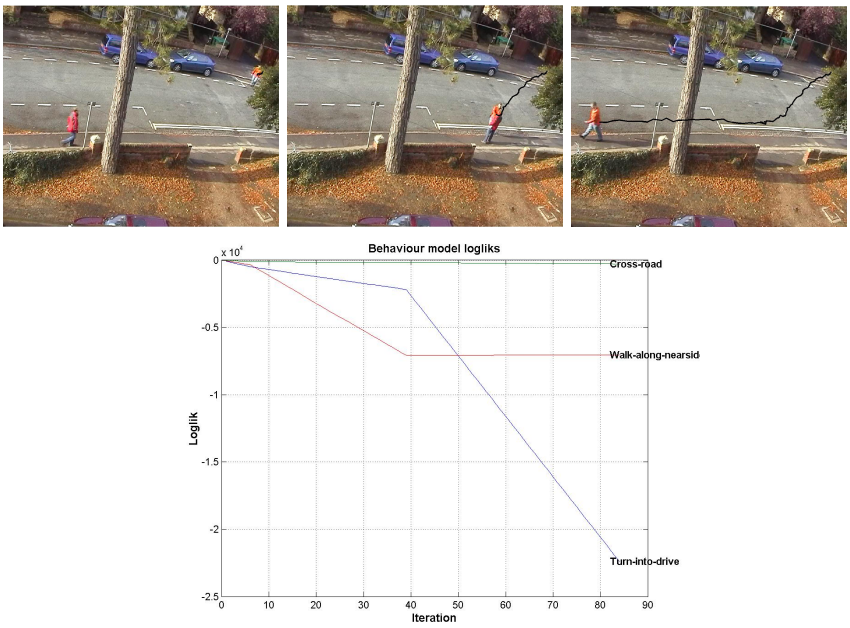


Figure 6: For the sequence in the top row we compute the likelihood of the extended behaviour models over the entire sequence. Note that each "iteration" is one computation step performed once for every 5 frames of video.

shown in figure 7.

5 Explaining interesting activity

Having automatically extracted human-readable descriptions of action, behaviour and gaze-direction for pedestrians in video, we are now in a position efficiently to encode a reasoning process to explain "interesting" activity. The overall process is based on predefined rules and is shown in figure 8. A set of "facts", derived from the application of the low-level activity recognition algorithms (described in section 4) to the input video stream, is maintained. This comprises all that is known about the agent's activity. For a particular scene, certain "trigger events" which require explanation are predefined, as are known rules about normal human behaviour for the scene. These can be encoded at a high-level only because we automatically derive qualitative human-readable descriptions of activity. The same reasoning process is used across video sequences from the same and very different application domains. (Although the trigger events and the rules require updating for different scenarios which can be done very efficiently.) As an illustrative example, in an urban surveillance scenario the event "move-to-road" is generated by a transition between the actions *walking-on-*

Figure 7: A single model associated with the *turning-into-drive* behaviour is used to classify the same behaviour but performed in different ways.

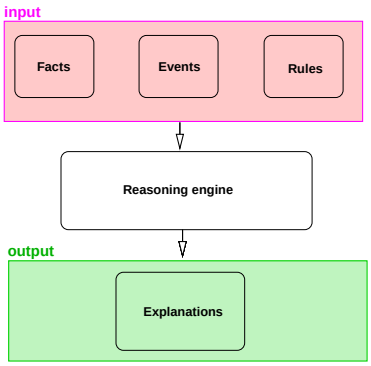


Figure 8: This diagram outlines the reasoning process we use for explaining activity in video. "Facts" are derived directly from video, "events" and "rules" are hand-coded for a particular scenario.

farside-pavement and *walking-on-road*. Intermediate events such as "meeting" or "ignoring" are inferred using rules which utilise all available information (including gaze direction). The hypothetical explanations for the activity are defined as follows:

1. IF the event "move-to-road" is followed by event "move-to-pavement" AND the current location is not the same as the location triggering the first event (i.e. the road is crossed) AND, subsequently, a meeting takes place THEN the explanation is that, "the agent crossed the road to meet the other agent",

2. IF a crossing of the road is observed NOT followed by an interaction THEN the explanation is that the agent crossed the road,

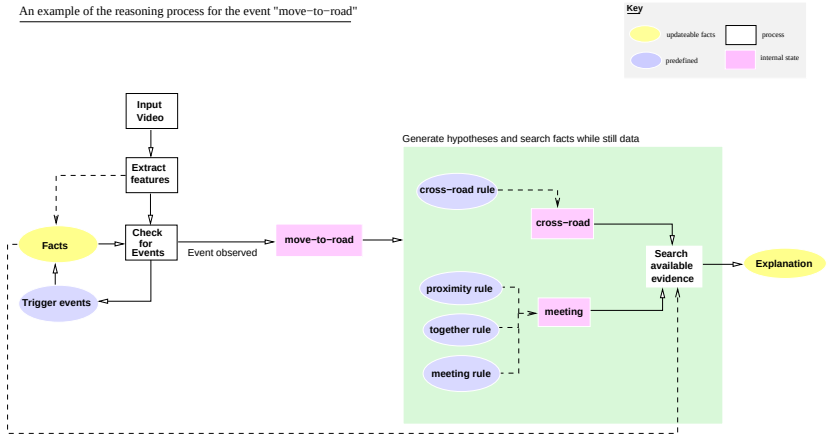


Figure 9: A schematic diagram of the reasoning process initiated when the event "move-to-road" is detected.

3. IF a "move-to-road" event is triggered AND subsequently a "move-to-pavement" event but back to the same pavement THEN no explanation is provided UNLESS another agent was in the near vicinity THEN the explanation is that it was necessary to avoid collision.

The extraction of qualitative descriptions of single-person activity results in very efficient encoding of rules about human behaviour in the scene extended over longer time periods. For example, we generate hypotheses for explaining events such as "stopping", "move-to-pavement" and "move-to-driveway". It can be seen that the rule-set is (a) general to all such urban scenes, (b) easily augmented (i.e. by quickly adding more rules). In figure 10, the output for two different situations, automatically generated by our system, is shown. The reasoning process is explicitly detailed for one example in figure 9. Exactly the same reasoning process and events-set is applied to the urban scene shown in figure 11. The rules are augmented with knowledge that the road may legitimately be crossed at the pedestrian crossing i.e. despite there being no evidence for a meeting, crossing at the lights is a plausible reason for the observed behaviour.

Finally, for interest and to demonstrate the generality of our reasoning process, in figure 12 we show an automatic explanation of traffic activity. In this case, the input activity description is limited (by comparison to the main results of this paper) but nonetheless demonstrates the utility of a rule-based reasoning system when an intermediate, qualitative estimation of low-level motion has been achieved.

5.1 Failure modes

The role of learning in an automatic reasoning system is not insignificant. While not explicitly implemented as a component of this system, we recognise that fail-

Explanation	Explanation
P2 move-to-road to Meet on ns-pavement	P1 move-to-road to Avoid P2 on ns-pavement
P2 move-to-pavement to Get-off-road	P1 move-to-pavement to Get-off-road

Figure 10: Explanations of interactions in an urban scene are automatically generated.

Commentary of activity	Explanation
NE-pavement, walking	Person move-to-road to Cross-road at N-ped-crossing
N-ped-crossing, walking	Person move-to-pavement to Get-off-road
N-ped-crossing, stopped	
N-ped-crossing, walking	
NW-pavement, walking	

Figure 11: The same rules and events set as used to generate the results in figure 10 is successfully used here in a different scene.

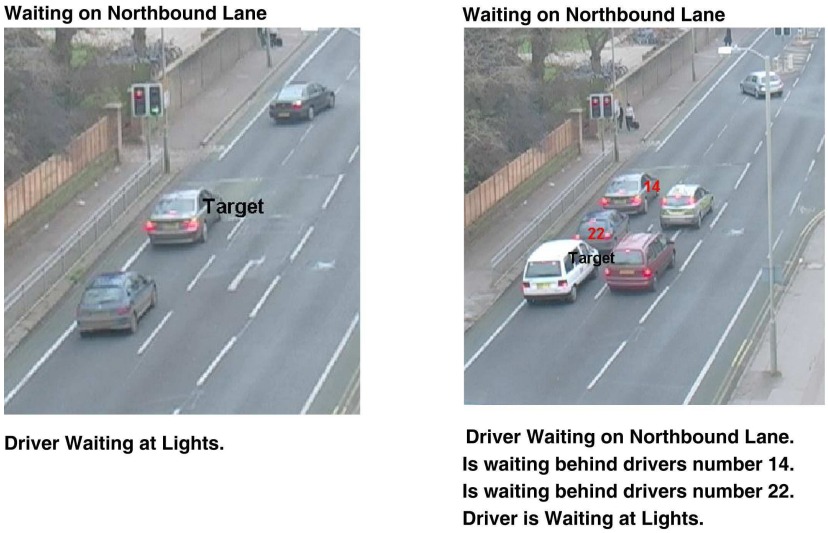


Figure 12: Resolution of potential anomalies i.e. why did the car stop? (*left*) and understanding of queues (*right*) is achieved using our method.

ure of any current implementation for a given scenario is either (a) an opportunity to learn or, (b) an opportunity to identify unusual/inexplicable behaviour. The latter could be used to prompt a surveillance analyst and would form a profitable strand of future research based on this work. In the case when no conclusion can be reached, the operator can be prompted to update the rule set to encompass the scenario encountered.

An example of an inexplicable event for the system presented in this paper is shown in figure 13, where a person is observed to walk on one pavement, the road and then return to the same pavement. Given the rule set as it existed at that time no explanation can be derived. It can be seen that this behaviour is, truly, inexplicable. However, were a car driving along the road (for example) an appropriate rule-fix might include knowledge of a pedestrian's desire to avoid traffic.

6 Conclusion

In this paper, we have presented a prototype system for generating high-level commentary on human activity in video and for reasoning about interesting events. We began by posing the question *What does an agent require to know in order to reason about a scene?* To answer this question we have exploited recent developments in Computer Vision with regard to action and behaviour recognition. The information we extracted was sufficient to enable not only the generation of accurate, human-readable commentary on surveillance video, but

Commentary on activity	Explanation
Nearside pavement, walking	Inexplicable
Road, walking	
Nearside pavement, walking	

Figure 13: This scenario is a failure mode for the system with the current rule set (as defined in the text). For a causal reasoning system failure is an opportunity to learn. In this case a true anomaly has occurred, although in other circumstances this behaviour could be exhibited when a car is present, for example.

also (and most significantly) causal explanations of interesting activity. This is the first demonstration of such a system which is (a) general for video sequences where the imaged person is low/medium resolution, and (b) complete, operating directly from the video stream to generate explanations of events, while utilising the "man-in-the-loop" and using complex visual features.

This research has considerable implications for the effective operation of CCTV surveillance. We have demonsrated how image and video processing techniques could be used within a semi-automatic process to help operators maintain global situational awareness of the entire scene when focussing on potentially interesting activity.

7 Future Work

We have described a semi-automatic approach to human behaviour understanding which aims to aid surveillance operators who can be overwhelmed with data, even from non-spurious targets. Future lines of research should, we believe, tackle the challenging problem of dealing with an increase in false alarm rates which will inevitably occur when automatic target identification and tracking is implemented for video surveillance, particularly when targets persist for a greater temporal period. As we stated above, the short life-span of objects in the video from our application domain, coupled with some operator intervention, mitigates against target appearance changes due to varying illumination or weather conditions.

Another pressing area for further development is to demonstrate fully probabilistic reasoning. For reasons of expediency, we have used the single most probable result from the vision components of our system to pass to the reasoning process, but we suggest that a Bayesian Network might equally well allow causal relationships to be inferred while retaining the benefits of probabilistic models (such as preventing committing to one decision too early in the reason-

ing chain). Pearl's work on Causality is likely to be relevant to this problem [13].

References

[1] E.L. Andrade, R.B. Fisher *Simulation of Crowd Problems for Computer Vision* First International Workshop on Crowd Simulation (V-CROWDS '05), Lausanne, Nov 2005

[2] J.L. Barron, D.J. Fleet, S.S. Beauchemin *Performance of Optical Flow Techniques* International Journal of Computer Vision 12:1 pp43-77, 1994.

[3] O. Boiman and M. Irani *Detecting Irregularities in Images and in Video* IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

[4] M.Brand, L.Birnbaum, P.Cooper *Sensible scenes: visual understanding of complex structures through causal analysis*, Proceedings, National Conference on Artificial Intelligence, Washington D.C., 1993.

[5] M. Brand and V. Kettnaker *Discovery and Segmentation of Actions in Video* IEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, August 2000.

[6] M.E. Bratman *Intention, Plans, and Practical Reason* CSLI Publications, Stanford University, ISBN (Paperback): 1575861925, 1988.

[7] D. Comaniciu and P. Meer *Mean Shift Analysis and Applications* Proceedings of the International Conference on Computer Vision-Volume 2, p.1197, September 20-25, 1999

[8] H. Dee and D. Hogg *Detecting Inexplicable Behaviour* Proceedings of the British Machine Vision Conference, 2004

[9] A.A. Efros, A. Berg, G. Mori and J. Malik *Recognising Action at a Distance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003

[10] M. Georgeff, B. Pell, M. Pollack, M. Tambe, M. Wooldridge *The Belief-Desire-Intention Model of Agency* Proceedings of the 5th International Workshop on Intelligent Agents V : Agent Theories, Architectures, and Languages (ATAL-98)

[11] E.T. Jaynes *Probability Theory: The Logic of Science* Cambridge, 2003, ISBN 0-521-59271-2

[12] B.Kuipers, Qualitative Reasoning, 1994 MIT Press, Cambridge, Massachusetts, USA.

[13] J. Pearl *Causality. Models, Reasoning and Inference* Cambridge University Press, 2000, ISBN 0 521 77362 8

[14] L.R. Rabiner *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proc IEEE, 77, 2, 257-285, 1989.

[15] M. Rigolli, *D.Phil. Thesis*, Department of Engineering Science, University of Oxford, 2006.

[16] N.M. Robertson and I.D. Reid *Behaviour understanding in video: a combined method* Proceedings of the International Conference on Computer Vision (ICCV), October 2005, Beijing, China

[17] N.M. Robertson and I.D. Reid *Estimating Gaze Direction from Low-Resolution Faces in Video* Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, May 2006

[18] S. Russel and P. Norvig *Artificial intelligence, a modern approach* Prentice-Hall, 1995

[19] B.J. Scholl *Innateness and (Bayesian) visual perception* In P. Carruthers, S. Laurence and S. Stich (Eds.), The innate mind: Structure and contents (pp. 34 - 52). Oxford University Press, 2005

[20] H. Sidenbladh M. Black, L. Sigal. *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking* European Conference on Computer Vision, Copenhagen, Denmark, June 2002.

[21] *Fuzzy Expert Systems and Fuzzy Reasoning William Siler* James J. Buckley ISBN: 0-471-38859-9, January 2005

[22] J.M. Siskind *Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic* Journal of Artificial Intelligence Research 15 (2001) 31-90

[23] P. Viola, M. Jones, D. Snow *Detecting Pedestrians using Patterns of Motion and Appearance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003 Pub. Morgan Kaufmann, Palo Alto, CA, USA, 1990.

[24] T. Xiang and S. Gong *Video behaviour profiling and abnormality detection without manual labelling* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005.

[25] L. Zelnik-Manor and M. Irani *Event-Based Video Analysis* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), December 2001

[26] H. Zhong, J. Shi and M. Visontai *Detecting Unusual Activity in Video* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004