A system for tracking and annotating illegally parked vehicles

Bogdan Vrusias, Dimitrios Makris, Ademola Popoola, Graeme Jones

A system for tracking and annotating illegally parked vehicles

Bogdan Vrusias[1], Dimitrios Makris[2], Ademola Popoola[1], Graeme Jones [2]
[1]*University of Surrey,* [2]*University of Kingston*
{b.vrusias, a.popoola}@surrey.ac.uk, {d.makris, g.jones}@kingston.ac.uk

Abstract

The large and still increasing number of vehicles on the roads has imposed the need for continuously monitoring the traffic behaviour, especially when it comes to illegally parked vehicles that disturb the normal flow of traffic. This paper presents an automatic method for identifying such events in CCTV video, by first tracking all the related objects and then annotating the events with appropriate keywords for storing and retrieval purposes. The method proposed makes use of a combination of video object tracking algorithms and techniques for capturing knowledge in keyword-based ontology structures. Starting from low-level visual information extracted from each video frame, high-level semantics such as moving objects are identified and classified. A probabilistic model, which takes its inputs from the visual modules, is used for identifying illegally parked vehicles. Finally, the keyword ontology, constructed automatically from expert descriptions, is linked to objects identified in the video. The results of the conducted experiments are encouraging.

1. Introduction

There has been a significant shift of interest of researchers in the area of visual surveillance, to the topic of event detection. This is justified by the maturing of motion detection [9] and motion tracking methods [10][11][12], but is also underpinned by the CCTV users who require systems that are able to detect specific type of events. For instance, the iLids project [15] identified 4 types of events that are of particular interest for the UK surveillance market.

Event detection can be broadly categorised in two categories: First, abnormal event detection aims to specify unusual events that are not expected to occur in the scene and may need the attention of the CCTV operator. A common approach is to measure the likelihood of an event using an appropriate probabilistic model (e.g. HMM in [13]) and define the outliers of this model to be the abnormal events.

Second, user-defined event detection aims to recognise specific events of interests for a given scene (e.g. "abandoning bag" or "intruding sterile zone"). Therefore, user's knowledge event is incorporated into to event detection mechanism, such a rule-based system or a Bayesian network [14].

However, user's knowledge needs to be encoded to the "system's language" and there is no space for users to communicate with the system using natural language. This may be important, because the events may described differently, based on the background, and experience of the CCTV operator.

Our work aims to bridge the gap between visual semantics that are detected by video processing algorithms and ontologies that are extracted by statistical natural language processing [2][16]. Our contribution is a system that processes and annotates videos for indexing and retrieval purposes. We focus on the problem of illegal parking, as presented by the iLids project.

The structure of the paper is as follows; Section 2 discusses the video processing methods to extract video semantics. Section 3 presents how a CCTV ontology is constructed using statistical natural language processing. The problem of linking visual semantics to ontologies is addressed in section 4. Section 5 focuses on the case study of "illegal" parking and demonstrates the results of our system. Finally, the paper is concluded in section 6.

2. Video Semantics

We identify three types of important semantics for the interpretation of CCTV videos: Actors, actions and static scene features. Combinations of those semantics are sufficient to describe the type of events that interest CCTV operators.

However, extracting these visual semantics from raw video is not straightforward. In our work, we firstly

estimate the blobs and the tracks of moving objects in the sequence, which then are used to estimate models of actors and areas. Finally, actor labels (pedestrian/vehicle) are attached to tracks and areas are marked according to their usage by the actors. Figure 1 shows the data hierarchy of this process.

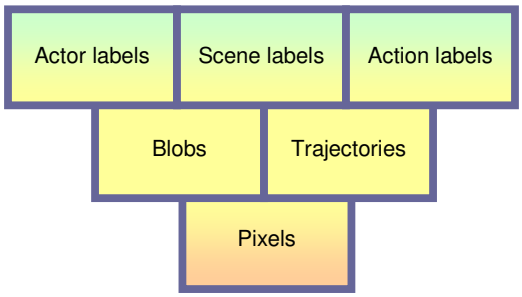

Figure 1. Data hierarchy

2.1. Object Tracking

We apply a motion detection algorithm [3] that is able to deal with illumination changes and shadows to separate foreground moving objects from background in individual fames. Then we use a motion tracking algorithm [4] to establish the temporal correspondence of these blobs, represented by a set of trajectories

Since colour and geometric features are important for the extraction of visual semantics, we ensure their invariance in time and 3D space respectively by colour [5] and geometric calibration [6]. Both types of calibration are automatic to avoid the practical issues of manual calibration, such as access to the physical place of the surveyed scene and the burden of human effort.

2.2. Object Recognition

Object Classification based on the geometric attributes of the detected blobs and the dynamics of their trajectories is used to distinguish between pedestrians and vehicles (Fig 2).

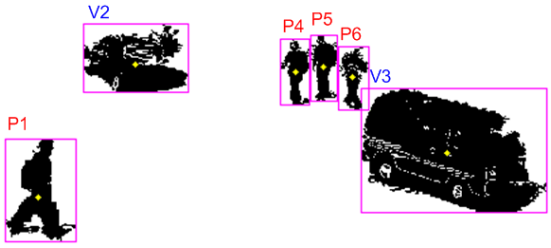

Figure 2. Pedestrians (P) and vehicles (V) attached to extracted blobs.

Our classification methodology uses an ensemble of classifiers that, currently, assigns objects to one of two classes: people & vehicles and, circumvents issues relating to feature selection [7]. Though primarily based upon Adaboost classifier, our methodology seeks to in-rich classifier performance through the determination of a metric's relative importance.

In order to store the output of the processed video, the information is encapsulated in predefined metadata files. The video metadata produced by the proposed method describes the tracked objects that can be captured in a video. The data stored for each object per frame includes: locations for bounding boxes, speed, bounding box size, object recognition label. Each object found in a video has its own XML file to and all of the separate XML files are summarized into a header XML file which contains a reference to all the individual files (Fig. 3).

Figure 3. An example of the header XML file and its relationship to an individual object XML file which describes a tracked object.

2.3. Scene Semantics

To determine activity-based scene features, we exploit the observed motion of the actors. We assume that activity-based scene features influence and/or enforce specific types of behaviour. For instance, roads constrain vehicles to move along specific lanes in a particular direction; similarly pavements constraints the pedestrian motion; gates, doors and regions at the image boundaries are related to entrance/exit events where actors will appear or disappear. Therefore,

accumulation of observations of the same behaviour in a specific location provides a clue for the existence of a correspondent activity-based feature [8].

3. Video Concept Relationship Structures

Most systems that capture knowledge lack of the ability to make this knowledge available to end users. Being able to annotate, index, and retrieve specific information without missing useful records is a challenging task. Video annotation in particular is one of the most challenging areas for information retrieval. To overcome this, a method for capturing the relationships between concepts used by experts in the process of video monitoring is described in this section. The method has three main stages: first, only the salient concept objects and actions are identified; second, the relationship between the concepts is determined and encoded in a domain specific ontology; and third, rules and reasoning are formed for the annotation purposes.

3.1. Identifying key objects and actions

The domain specific multimedia ontology is intended for documenting the semantics of video scenes that cannot easily be captured through video analysis techniques alone. Statistical text analysis is used to capture this ontology from expert descriptions of videos. Other existing ontology structures are used in order to complement the one built. The ontologies used here are acquired from two sources: the WordNet lexical database and the Police Information Technology Organisation (PITO), now the National Policing Improvement Agency (NPIA), terminology.

The components built and used in this process integrate with Sheffield's GATE system. These components build on existing GATE plug-ins from ANNIE, for preliminary NLP tasks of such as POS tagging and sentence splitting. OWL is used as the syntax for representing the ontology which captures the weight of the co-occurrence relationship between objects and actions concepts found in the videos [1]. The relationship weight is calculated using a function of frequency and distance between terms, with the purpose to later determine the strongest relationships between objects and actions. This information is the further used to associate the identified video objects with concepts in the ontology, for potential use in video search expansion.

The pipeline for these resources are (Figure 4):

1. Linguistic Concept Identification (2.1)
2. Statistical Concept Identification (2.2)
3. WordNet Verification (2.3)
4. PITO Terminology Analysis (2.4)
5. Event Mapping (2.5)

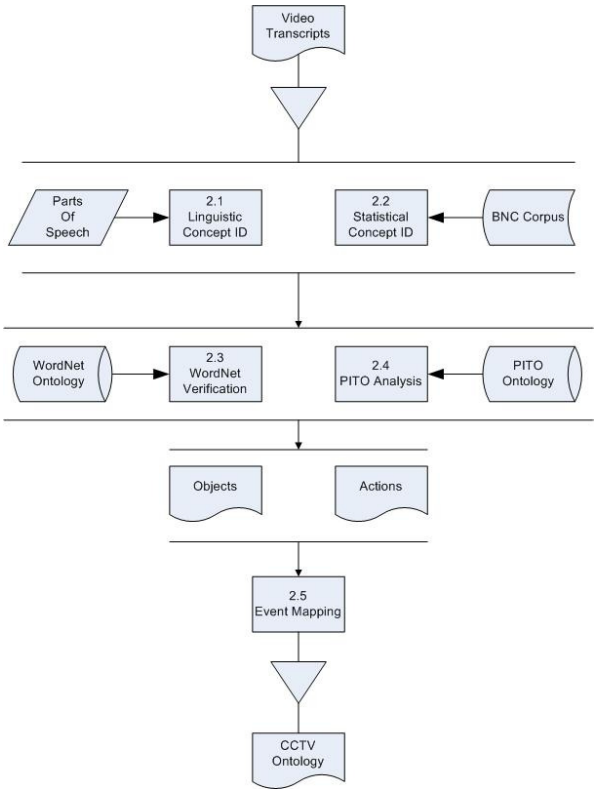

Figure 4. Pipeline for the automatic generation of the term-based CCTV ontology [1].

In order to identify objects and actions in the text descriptions of videos (transcripts), the GATE ANNIE tokeniser and Part of Speech (PoS) tagger are used to identify all the nouns and verbs, and to provide a basis for the identification of compound nouns according to specified patterns of PoS annotations. Some basic stemming does occur for the initial assessment of potential object and action concepts. At this point no frequency or other statistical information is considered.

The initial linguistic identification of concepts is supplemented by a statistical approach to identify other potential concepts that may have been incorrectly classified by the PoS tagger. Transcripts undergo statistical analysis to determine salience using frequency and weirdness information to provide statistical evidence for concepts and to act as confirmation for the linguistic extraction [1]. Results are classified as either objects or actions via the WordNet Verification (2.3) process. The British National Corpus (BNC) is then used as a reference corpus, and thresholds are adjusted by modifying

parameters for the distributions. Similar approaches that process text, available with multimedia, and extract terms using combinations of word frequency, tf*idf and entropy, and application of stemming [17], where resulting keywords are used in the manual construction of an ontology.

Information and procedures defined by the National Policing Improvement Agency (NPIA), formerly known as the Police Information Technology Organisation (PITO), are also used as a filtering stage in the process of building the video ontology. PITO/NPIA specified a terminology for describing events or incidents. PITO provide data definitions which help sub-divide information into simpler data elements. The PITO/NPIA elements were transformed into an ontology by re-interpreting the category and sub-category information as classes and sub-classes. The PITO/NPIA ontology is used for filtering purposes when classifying objects and actions in the generated ontology. This information is used to associate keywords for more efficient relationship analysis in Event Mapping (Figure 4, 2.5).

3.2. Building a video ontology

Several knowledge structures for holding data have been used in the literature; the ontology data structure is used here as a way for storing and retrieving relationships between video concepts of objects and actions. Domain ontologies are useful for sharing knowledge between more than one application on the web, and they are structured in such a way that allows machine-based communication. They are often used in e-commerce to allow machine-based communication between buyer and seller. They are also used in some search engines to allow searching of words which are semantically similar but syntactically different therefore enhancing the search results, this method is sometimes known as query expansion, and is adopted for this paper too.

With objects and actions identified through the previous processes and filtering thereof, we analyze connections between the objects and actions to produce two types of event, 'action-object' or AO event and 'agent-action-recipient' or AAR events. An AO Event associates one object and one action. An AAR Event represents the relationship between two objects and the action that connects them. We calculate collocation distances between words connected in events to ascertain dominant patterns. The connections are used to create instances of the AO Event and AAR Event classes in the ontology.

The REVEAL project makes use of OWL DL and the RDF/XML syntax to produce a representation of the objects and actions contained within a surveillance video and the events linking them together.

3.3. Expanding the video ontology

The videos that will be searched by the system contain *Objects*, *Actions*, *ActionEvents* and *ObjectEvents* contained within the ontology. This ontology was produced from descriptions of videos given manually by several different CCTV experts watching them. The process of creating the ontology looks at the descriptions and locates object related and action related words, and the distance between them:

Object: a noun that relates to an object that was described in the video, such as 'person' or 'car'.

Action: a verb that relates to an action that was described in the video, such as 'cross' or 'drive'.

ActionEvent: this relates to an event described in the video in the form action_object, such as 'cross_person'.

ObjectEvent: this relates to an event described in the video in the form object_object_action, such as 'car_person_drive'.

An exemplar output of the system is shown here:

```
<rdf:Description rdf:about="car_stop_road">
  <agent rdf:resource="car"/>
  <action rdf:resource="stop"/>
  <recepient rdf:resource="road"/>
  <freq>7</freq><dist>3.5</dist>
</rdf:Description>
```

The query expansion that was performed uses the WordNet ontology and is created to supply synonyms, hyponyms and meronyms as the search criteria specified, and use these words along with the original criteria when searching the generated video ontology. Different levels of expansion for each of the search words, supplied by the user, are available, and also narrowing the expansion performed. Although this sometimes led to some inaccurate results being returned.

4. Linking video semantics to ontologies

The combination of different modalities of the same information can produce high-performance retrieval systems. The *semantic gap*, is the difference between ambiguous formulation of contextual knowledge in a powerful language e.g. natural language and its sound, reproducible and computational representation in a formal language e.g. programming language [2][16].

A Bayesian Network is used in this study as a 'concept detector'. Features like Elapsed Time, Distance to Stop Zone, obtained from experts in a workshop conducted during the REVEAL project, were used to construct the Bayesian network. As suggested by [18], the network structure was designed by incorporating so-called synthetic nodes in order to reduce the number of parents of each child node. In order to reduce bias during probability elicitation from experts, we followed the method used in [19], where probabilities were obtained using text fragments and a scale with both numerical and verbal anchors.

The system relies on the use of keywords or tags generated from the low level object detector/recognition (section 2.2) and the concept detector (Bayesian network). However, two major problems are associated with "plain" keywords: synonymy and polyzemy, which affect the performance of semantic retrieval. The concept synonymy is used to describe the fact that there are many ways to refer to the same object, while polyzemy is referred to the fact that most words have more than one distinct meaning [20]. We have taken advantage of the lexical relationships – synonymy, hyponymy and hypernymy - in WordNet to carry out tag expansion. The pipeline for these resources is shown in Figure 5.

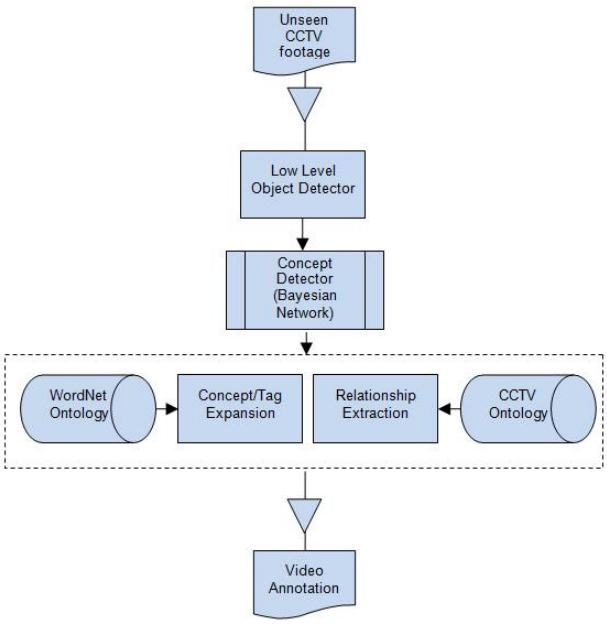

Figure 5: Pipeline for the automatic annotation of CCTV footage

5. Case Study: tracking "illegally" parked vehicles

The dataset was supplied by the REVEAL project and consisted of 12 videos, manual descriptions of those videos, an ontology describing the objects, actions and events occurring within all 12 videos and video metadata supplying the bounding box locations for objects identified within the video. The descriptions for the videos were produced by 6 different people each commenting on the same 12 videos. Videos are split into 20 second periods and the 6 descriptions, from the 6 different people, for that 20 second time period are merged into one file.

The videos are segmented into 20 second periods, each of which has its own description and for each query used in the evaluation the segments of video that are expected to be returned are marked manually. Then, using the expected results and the actual results from the system precision and recall are calculated. Due to the fact that the expected results rely on the integrity of the video descriptions given by people it is feasible that a description could be wrong, however it is assumed for this evaluation that the descriptions are reliable.

Figure 6: Scene from a video

"On the left hand side is a parked white van with its rear doors open with a small dark saloon parked just to the rear of it with a gap of several feet between. There is a car parked on the right hand side of the near side and pedestrians walking towards the camera."

"The white van on the far left of the screen, the door is now open with the lift platform by itself, or the van sorry left by itself with the doors open. Pedestrians walking up and down, cars going down, there does not seem to be too much happening."

Figure 7: Two examples from six transcripts describing the same scene as featured in figure 6.

The main focus is placed in identifying 'illegally' parked vehicles, so the clips are described in terms of the presence of 'interesting events', which are vehicles initially in motion that subsequently become stationary i.e. have a velocity = 0. Within the set of 'interesting' events, a subset of that, 'illegally' parked vehicles, is identified. Conditions for a vehicle to be classified as being 'illegally' parked, it had to be identified as (i) an object previously in motion (ii) classified as a 'vehicle' (iii) become stationary for more than one minute (as defined by the NPIA for the iLids dataset) and (iv) be parked in a 'no stop' location. A total of 41 interesting events were identified in the 12 clips used for this case study, with 18 of these events - almost half – further classified as illegally parked vehicles (Table 1).

Table 1. 'Interesting' events & 'illegally' parked vehicles in 12 clips

Clip No.	'Interesting' Event Count	'Illegally' Parked Vehicle Count
1	4	2
2	3	1
3	7	3
4	7	3
5	4	1
6	1	1
7	3	0
8	2	1
9	1	1
10	2	0
11	4	2
12	3	3
Totals	**41**	**18**

In each clip, the low level object detector automatically identify and classify moving objects, using initial tags: 'vehicles' and 'person', as described in section 2.2 (Figure 8). Since the aim of this case study is to annotate 'illegally' parked vehicles, we focus on objects classified as 'vehicles'. When a previously moving 'vehicle' object stops, the object is selected as 'interesting', and its properties, in relation to other objects present in the same timeframe, are collated. Features like Elapsed Time, Distance to Stop Zone, obtained from experts in a workshop conducted during the REVEAL project, were used to construct a Bayesian network. The data obtained from the video footage are used in the Bayesian network to classify the stopped vehicle as being 'illegally' parked i.e. 'abandoned' or not. Thus, the Bayesian network is used as a 'concept detector', the concept being 'abandoned' or 'illegally' parked vehicles.

Figure 8. Initial tags/labels from Low Level Object Detector.

Once a 'vehicle' object is identified or labelled as 'abandoned' or illegally parked, we use the lexical relationships in WordNet to perform tag or label expansion in order to generate additional tags or labels. For example, the 'vehicle' tag generates hyponym expansions such as 'motor vehicle' and 'car'. The initial tag and associated expansions are then used to query the CCTV ontology, to extract existing, statistically significant relationships that are linked to these tags. The relationships between concepts determined through relationship extraction module are detailed in Table 2, showing the 'agent-action-recipient' or AAR events.

The initial and expanded tags, as well as associated relationships are then used to 'annotate' the video footage (Figures 9-10.)

Table 2: The top 10 identified AAR events from the transcripts

AAR Event	Count
Car-Come-Road	20
Person-Cross-Road	16
Pedestrian-Cross-Road	11
Car-Go-Road	9
People-Cross-Road	9
People-Have-Discussion	7
Pedestrian-Walk-Side	6
Car-Come-Side	5
People-Walk-Street	4
Somebody-Cross-Street	4

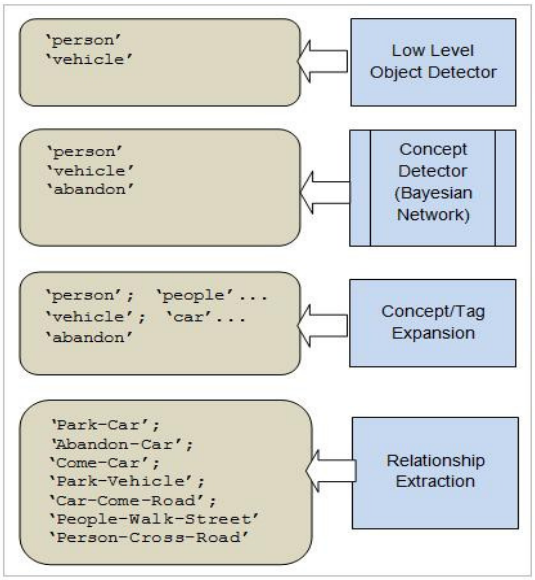

Figure 9. Exemplar annotations produced by each module of the system

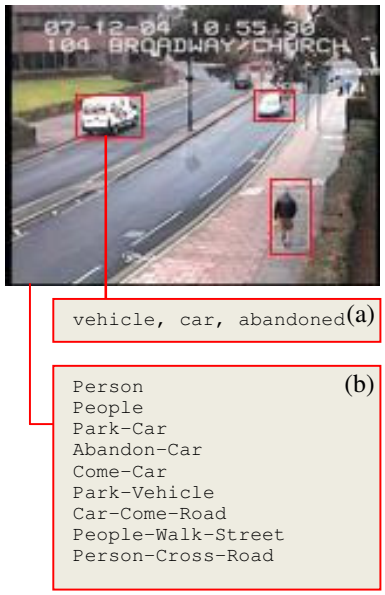

Figure 10 Automatically identified keywords related to the object in the bounding box (a), and then generated annotation for the *frameshot* (b).

6. Conclusions

A system for automatically annotating "illegally" parked vehicles is described. The method combines video processing for identifying blobs and consequently conceptual entities in videos, together with statistical text analysis methods for building knowledge structures that capture relationships between video content concepts. In order to extract high level semantic annotations from the videos, the visual and text extracted semantics are combined via a Bayesian network classifier. The system successfully identified the "illegally" parked vehicles present in the videos. Although the case study focused on a particular scenario, the method followed can potentially be used to identify other scenarios such as people running, traffic congestions, overtaking, or unusual behaviour in general, which can be of great use in traffic control or public safety.

7. Acknowledgement

This work was supported EPSRC sponsored REVEAL project under Grant No.GR/S98450/01, GR/S98443/01 jointly undertaken by the University of Surrey and Kingston University. We are grateful to our project partners and expert end-users, John Armstrong from Surrey Police, the UK National Policing Improvement Agency (NPIA), and Home Office - Scientific Development Branch (HOSDB). Special thanks to Neil Newbold for building the domain ontology system and David Williams for building the video retrieval system.

8. References

[1] N. Newbold, B. Vrusias and L. Gillam, "Lexical Ontology Extraction Using Terminology Analysis: Automating Video Annotation", *The 6th Int. Conf. on Language Resources & Evaluation (LREC)*, Morocco, 2008.

[2] B Vrusias, D. Makris, J.R. Renno, N. Newbold, K. Ahmad and G.A. Jones, "A Framework for Ontology Enriched Semantic Annotation of CCTV Video", *Int. Workshop on Image Analysis for Multimedia Interactive Services*, Santorini, Greece, 2007.

[3] J.R. Renno, N. Lazarevic-McManus, D. Makris, G.A. Jones, "Evaluating Motion Detection Algorithms: Issues and Results", Sixth IEEE International Workshop on Visual Surveillance, May 13, Graz, Austria, pp. 97-104., 2006.#

[4] M. Xu, T.J. Ellis, "Partial observation vs. blind tracking through occlusion", British Machine Vision Conference, BMVA, September, Cardiff, pp. 777-786, 2002.

[5] J.R. Renno, D. Makris, T.J. Ellis, G.A. Jones, "Application and Evaluation of Colour Constancy in Visual Surveillance", Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation (VS-PETS 2005), China, Beijing, 2005.

[6] J.R. Renno, J. Orwell, G.A. Jones, "Learning Surveillance Tracking Models for the Self-Calibrated Ground Plane", British Machine Vision Conference, September, Cardiff, pp. 607-616, 2002.

[7] J.R. Renno, D. Makris, G.A. Jones, "Object Classification In Visual Surveillance Using Adaboost", Seventh International Workshop on Visual Surveillance, Friday June 22nd, Minneapolis, 2007.

[8] D. Makris, T.J. Ellis, J. Black, "Learning Scene Semantics", ECOVISION 2004 Early Cognitive Vision Workshop, May, Isle of Skye, Scotland, UK, 2004.

[9] C. Stauffer, W.E.L Grimson, "Adaptive Background Mixture Models for Real-Time Tracking". Int. Conf. on Computer Vision and Pattern Recognition, CVPR99, Fort Colins, USA, 1999.

[10] I. Haritaoglu, D. Harwood, L.S. Davis, W^4: Real-Time Ssurveillance of People and their Activities, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8), 809-830, 2000

[11] Michael Isard and Andrew Blake, CONDENSATION -- conditional density propagation for visual tracking, International. Journal on. Computer Vision, 29, 1, 5--28, (1998).

[12] D. Comaniciu, P. Meer: Mean Shift: A Robust Approach toward Feature Space Analysis, IEEE Trans. Pattern Analysis Machine Intell., Vol. 24, No. 5, 603-619, 2002

[13] A. Galata, N. Johnson, D. Hogg, Learning variable-length Markov models of behavior, Computer Vision and Image Understanding, 81 (3) (2001) 398-413.

[14] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods, Computer Vision and Image Understanding, 96(2), pp 129-162, 2004.

[15] Home Office Scientific Development Branch
Imagery library for intelligent detection systems (i-LIDS), http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/, [Last accessed: May 2008]

[16] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser and C. J. Sandom, "Bridging the semantic gap in multimedia information retrieval - top-down and bottom-up approaches," *In Proc. 3rd European Semantic Web Conference*, Budva, 2006.

[17] A. Jaimes and J. R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," *In Proc. of IEEE Int'l Conference on Multimedia & Expo*, Vol. 1, pp. 781-4, 2003.

[18] Martin Neil , Norman Fenton , Lars Nielson, Building large-scale Bayesian networks, The Knowledge Engineering Review, v.15 n.3, p.257-284, September 2000

[19] L. C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B. G. Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. Artificial Intelligence in Medicine25, pp. 123 - 148, 2002.

[20] Changbo Yang, Ming Dong, Farshad Fotouhi, "Learning the Semantics in Image Retrieval - A Natural Language Processing Approach," cvprw, p. 137, 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 9, 2004.