A Framework for Ontology Enriched Semantic Annotation of CCTV Video

Bogdan Vrusias,
University of Surrey, UK
b.vrusias@surrey.ac.uk,

Dimitrios Makris,
Kingston University, UK
d.makris@kingston.ac.uk,

John-Paul Renno,
Kingston University, UK
j.r.renno@kingston.ac.uk,

Neil Newbold,
University of Surrey, UK
n.newbold@surrey.ac.uk,

Khurshid Ahmad,
Trinity College, Ireland
khurshid.ahmad@cs.tcd.ie,

Graeme Jones
Kingston University, UK
g.jones@kingston.ac.uk

Abstract

This paper deals with the problem of semantic transcoding of CCTV video footage. A framework is proposed that combines Computer Vision algorithms that extract visual semantics, together with Natural Language Processing that automatically builds the domain ontology from unstructured text annotations. The final aim is a system that will link the visual and text semantics in order to routinely annotate video sequences with the appropriate keywords of the domain experts' terminology.

1. Introduction

CCTV systems have an increasing role as means of surveying public places. Collected surveillance video data is normally used to resolve crimes and used as evidence in courts. However, the large numbers of installed CCTV cameras that operate continuously produce enormous video datasets, in which users need to search to access video shots that are semantically interesting. Therefore, video footages need *semantic transcoding* [1], i.e. indexing of video shots with the appropriate semantic labels. Ideally, the semantic content should be described in the specialised domain language that is used by the operators of the system.

Considerable research effort has been invested in the area of video analysis of CCTV footage especially in areas of low-level video processing, such as motion detection and motion tracking [2]. However, Video-based semantic transcoding is required to identify the interesting visual semantics and appropriately tag the video shots. Research in the area of Natural Language Processing indicates that the semantics used to describe CCTV video footage fall into three main categories: *agent*, *action* and *recipient*. Other important semantic properties captured in video footages include *location*, *time*, and *direction* [1][3][4]. Agents and recipients are detected by motion segmentation techniques and labelled by an object classification algorithm [5]. Event detection methods use a-priori scene-based knowledge to identify interesting events by analysing the motion histories of agents and recipients [6]. As yet little research has explored the detection of the static features of the scene (i.e. locations such as routes or exit and entry zones). In [7] a colour-attention mechanism is used to segment size and position information to label static features, while [15] uses a motion-attention mechanism to determine activity-based static scene features by accumulating activity observations.

There is limited research on text-based approaches for constructing the video ontology. Most common methods of constructing the video ontology are derived directly from the video processing with limited success. Ontology abstractions, on the other hand, depend less on low-level video features and more on the domain ontology [1]. Terminology-based semantic transcoding helps identifying the key concepts within a video scene that could reduce the burden on the visual processing.

Researchers have explored the application of natural language process (NLP) methodologies in which objects detected in a video scene are treated as *concept terms* that, based on their frequency, can define the semantics of the scene [1][7]. What is missing, though, is what might be called "expert knowledge of the domain" which lies embedded within the experiences of the human expert. However, manually building such a knowledge base is a laborious and expensive task.

One important development is the emergence of markup standards for videos. The key scripting languages include Web Ontology Language (OWL[1]),

[1]http://www.w3c.org/2004/owl

Eight International Workshop on Image Analysis for Multimedia Interactive Services(WIAMIS'07)
0-7695-2818-X/07 $20.00 © 2007 IEEE

Resource Description Framework (RDF[2]), and XML within the MPEG-7[3]. Another scripting language, more specific to video, is the video event representation language (VERL), and its 'companion' VEML, that have been recently proposed as 'an ontology framework for representing video events' [4].

We have discussed elsewhere how terms can be extracted automatically from a corpus of documents belonging to a specialist domain [8], organised into a candidate ontology [9][10], and subsequently be used to annotate and retrieve crime scene images [11]. We have discovered that experts use a common language and keywords to describe video evidence. However, research that combines the two technologies of Computer Vision and Natural Language Processing is very limited. We propose a framework that bridges the gap between the visual semantics extracted from video footage and the ontology that is built from text annotations. A system architecture is presented that will automatically annotate unseen videos with appropriate keywords. Such an approach is also important from a cognitive point of view as the proposed system mimics the human learning ability to combine their visual experiences with contemporaneous verbal discussions.

2. Video Annotation Framework

Several attempts have been published on bridging this *semantic gap* between visual and terminology semantics [3][6][7]. However, most such attempts refer to annotating images rather than video. The additional dimension of time, therefore, must also be embedded within the ontology. Ontology for the CCTV domain is specific because operators have knowledge of objects and events that are important for the domain and ignoring other "common" objects or events.

We have constructed a framework specific for the CCTV domain, which combines the visual and terminology semantics to produce a *surveillance metadata model*. The high level semantics extracted from videos are linked to concepts extracted from video descriptions (fig. 1). The framework supports the linkage of video detected objects with concepts from the ontology. In this manner, the video can be later queried with keywords related with these concepts.

A data hierarchy scheme is proposed to facilitate the bridging of gap between the input (video raw data) and the expected output (textual description) of the system. (fig. 2). Raw data consists of pixels in consecutive frames. Video Processing will segment video sequences into moving objects, described by blobs and trajectories and Machine Learning methods will extract the visual semantics: actors (agents), locations or static scene objects (recipients), and events (section 5). NLP of human-derived annotations will automatically build the CCTV-domain ontology and the Visual Evidence Thesaurus (Section 4). The visual semantics will be annotated by the keywords of the CCTV ontology and textual summaries of video shots will be provided.

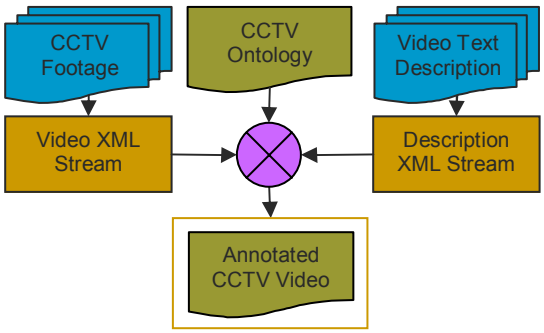

Fig. 1. Overview of the framework showing the process of video annotation

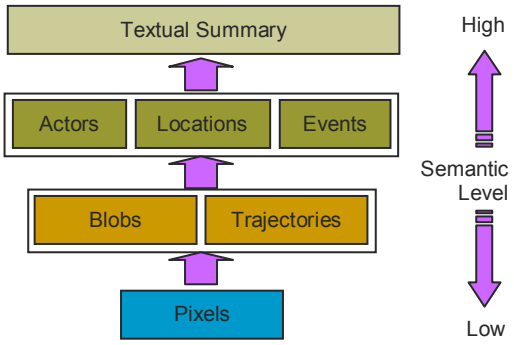

Fig. 2. Data hierarchy representing different levels of information extracted from video.

3. Surveillance Metadata Model

The main purpose for automatically constructing CCTV ontology from text is to capture the semantics of the video scene that are not directly available from video analysis. This ontology can be used to enhance the video information and help with the semantic transcoding. There are three important features in video transcriptions: the moving objects, the passive scene objects, and the actions performed by the *moving object(s)* either as an *interaction* with the *passive object(s)* or other moving object(s). These objects can be identified directly from within video descriptions, or from hand-made domain ontologies such as the PITO (Police Information Technology Organisation[4])

[2]http://www.w3.org/RDF
[3]http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
[4] http://www.pito.org.uk

0-7695-2818-X/07 $20.00 © 2007 IEEE

ontology. We have therefore organised and run workshops to collect descriptions of videos in a free-speech format, given by CCTV operators and police officers. The description were later transcribed into free-text format and added to the domain corpus, ready to be processed. Finally, we encoded the PITO ontology from a raw text format to a structured OWL format, and added that to the main framework as well.

We decided to use OWL as the scripting language for constructing the ontology, and Protégé[5] was used for visualising the semantic relationships. We also decided to use the Jena OWL API which is a Java framework for building Semantic Web applications and provides a programmatic environment for OWL. The Kazuki extension to Jena was also used for handling concept instances within the generated ontology. Finally, we extended the well-established GATE[6] text analyser system, by adding additional support for identifying, highlighting, and exporting the conceptual relationships identified within texts.

From our analysis we also identified some common expert description patterns. Usually experts first identify and describe the most important *objects*, then elaborate the *event*, and finally describe the *location*. This information is also captured within the ontology.

4. Ontology and Thesaurus Extraction

To generate the text-based ontology we initially tokenise the video descriptions and each token is tagged with the appropriate PoS (Part of Speech) label. Tokens are then filtered based on their weirdness value [8] and their presence in the PITO ontology, and then separated into two lists: the object related words (actors or recipients), and the action related words. The action words are used for the collocation process [9]. For each action word we identify the object word(s) that it is related to, and for each object word we identify the other object word(s) that it is related to via the action word (triplets). We form a semantic network where we describe action and object words together with the frequency of occurrence and the average distance between them. We also automatically create links to other ontologies such us the WordNet and the PITO ontology. Finally using OWL as the scripting language we automatically capture the semantics of the video descriptions into CCTV ontology (fig. 3).

The ontology captures the *weight* of the co-occurrence between objects and actions. The weight is measured in terms of *frequency* and *distance* between the co-occurred terms. This measure is used for calculating strongest relationships between objects and events, which is then used for retrieving video scenes.

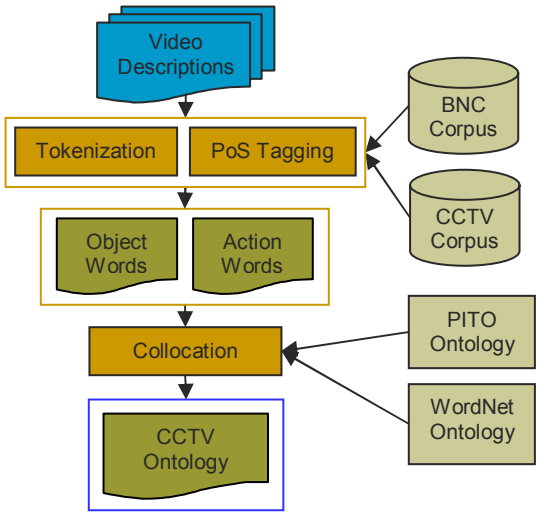

Fig. 3. Automatic generation of term-based video ontology

5. Extracting Visual Semantics

Visual Semantics are extracted by a series of processes on different levels of the data hierarchy of fig. 2. We apply motion detection to extract blobs in individual frames and then a motion tracking algorithm [12] 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 [13] and geometric calibration [14]. Both types of calibration are automatic to avoid the issues of manual calibration, such as access to the physical place of the surveyed scene and the burden of human effort.

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. 4) [5].

One of the key novelties of our approach is the extraction of static scene features. We exploit the observed motion of the actors to determine activity-based scene features. It is assumed 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; gates and doors are related to entrance/exit events where actors will appear or disappear; and bus stops indicate where pedestrians wait for the bus. Therefore, the accumulation of observations of the same behaviour in a specific location provides a clue for the existence of a correspondent activity-based feature [15].

[5] http://protege.stanford.edu
[6] http://gate.ac.uk

0-7695-2818-X/07 $20.00 © 2007 IEEE

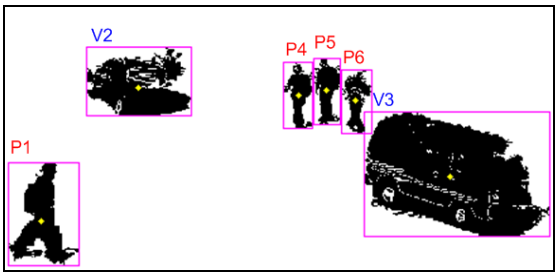

Fig. 4. Detected blobs classified as pedestrians (P) or vehicles (V).

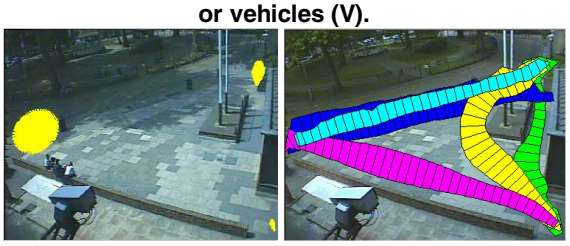

Fig. 5. Automatically derived entry/exit zones and routes

6. Conclusion

We described a framework that links the technologies of Computer Vision and Natural Language Processing in the application domain of Visual Surveillance. The resulting schema dynamically merges the visual and textual ontology together, in order to produce systems that can automatically annotate unseen video footage with appropriate keywords that have been identified in human annotations of other videos. Such an approach is important not only for its enormous application value which is the facilitation of automatic retrieval of CCTV video shots, but also because it attempts to simulate the cognitive ability of humans to learn by combining visual experiences with verbal information. Furthermore, we have shown how domain experts use a specialist language when describing video scenes and how the semantics of the language are captured within an ontology structure automatically. The fusion of visual and terminology based semantics has lead to a comprehensive model that supports semantic transcoding within a video scene which could potentially reduce the burden on the visual processing.

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 PITO, and HOSDB.

8. References

[1] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic segmentation and description for video transcoding," *Proc. of IEEE Conf. on Multimedia & Expo*, vol. 3, pp. 597, 2003.

[2] W. Hu, T. Tan, L. Wang, S. Maybank, "A survey on visual surveillance of object motion and behaviors", IEEE Trans. on Systems, Man, and Cybernetics, Part C. Vol. 34, no. 3, pp. 334-352. Aug. 2004.

[3] M. Bertini, R. Cucchiara, A. Del Bimbo, and A. Prati, "An Integrated Framework for Semantic Annotation and Transcoding," *Multimedia Tools and Applications*, vol. 26, no. 3, pp. 345-363, 2005.

[4] A.R.J. Francois, R. Nevatia, J. Hobbs, and R.C. Bolles. "VERL: An Ontology Framework for Representing and Annotating Video Events," *IEEE MultiMedia*, vol. 12, no. 4, pp. 76-86, 2005.

[5] P. Remagnino, G.A. Jones, "Classifying Surveillance Events from Attributes and Behaviour", British Machine Vision Conf., pp. 685-694, Manchester, UK, Sept. 2001.

[6] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 23, no. 8, pp. 873–889, 2001.

[7] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris,V.K. Papastathis, and M.G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 10, 2005.

[8] K. Ahmad and M. A. Rogers, "Corpus-Based Terminology Extraction", *Handbook of Terminology Management*, vol. 2, pp. 725-760, 2000.

[9] K. Ahmad and L. Gillam. "Automatic Ontology Extraction from Unstructured Texts", In (Eds.) R. Meersman and Z. Tari. *On the Move to Meaningful Internet Systems - OTM*, Confederated Int. Conf., CoopIS, DOA, and ODBASE 2005, Agia Napa, Part II, pp. 1330 – 1346, 2005.

[10] K. Ahmad, M. Tariq, B. Vrusias and C. Handy, "Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains", 25th EU Conf. on Info. Retrieval Research, Pisa, pp. 502-510, 2003.

[11] B. Vrusias, M. Tariq, and L. Gillam, "Scene of Crime Information System: Playing at St. Andrews," Comparative Evaluation of Multilingual Info. Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, LNCS, vol. 3237, pp. 631-645, 2005.

[12] M. Xu and T. Ellis "Partial observation vs. blind tracking through occlusion" *British Machine Vision Conf.*, Cardiff, pp. 777-786, Sept. 2002.

[13] J.R. Renno, D. Makris, T.J. Ellis and G.A. Jones, "Application and Evaluation of Colour Constancy in Visual Surveillance", *IEEE Int. Workshop on Visual Surveillance and Performance Evaluation*, China, Beijing, 2005.

[14] J.R. Renno, J. Orwell and G.A. Jones, "Learning Surveillance Tracking Models for the Self-Calibrated Ground Plane", *British Machine Vision Conf*, Sept. 2002.

[15] D. Makris and T. Ellis, "Learning Semantic Scene Models from Observing Activity in Visual Surveillance," *IEEE Trans. on Systems Man and Cybernetics - Part B*, vol. 35, no. 3, pp. 397-408, June 2005.