Lexical Ontology Extraction Using Terminology Analysis: Automating Video Annotation

Neil Newbold, Bogdan Vrusias, Lee Gillam
University of Surrey
Surrey GU2 7XH, UK
E-mail: n.newbold@surrey.ac.uk, b.vrusias@surrey.ac.uk, l.gillam@surrey.ac.uk

Abstract

The majority of work described in this paper was conducted as part of the Recovering Evidence from Video by fusing Video Evidence Thesaurus and Video MetaData (REVEAL) project, sponsored by the UK's Engineering and Physical Sciences Research Council (EPSRC). REVEAL is concerned with reducing the time-consuming, yet essential, tasks undertaken by UK Police Officers when dealing with terascale collections of video related to crime-scenes. The project is working towards technologies which will archive video that has been annotated automatically based on prior annotations of similar content, enabling rapid access to CCTV archives and providing capabilities for automatic video summarisation. This involves considerations of semantic annotation relating, amongst other things, to content and to temporal reasoning. In this paper, we describe the ontology extraction components of the system in development, and its use in REVEAL for automatically populating a CCTV ontology from analysis of expert transcripts of the video footage.

1. Introduction

The ever-expanding web, and the increased prevalence of multimedia data, places a substantial burden on users in understanding and rapidly processing large volumes of information and working through various vocabularies and classifications as often evidenced through ontologies and folksonomies. Semantic Web technologies promise to be able to analyse and automatically annotate and relate many different kinds of multimedia documents, including video, image, music and sound, text, and various fragments of these. It is not only in the Semantic Web where automatic annotation could be beneficial: automatic annotation can be useful in the ever-present fight against crime. Closed Circuit Television (CCTV) is an increasingly popular way of enabling remote monitoring and policing of specific (visual) scenes such as ATMs, retail outlets, football (soccer) matches, in airports, in railway stations, in car parks, and so on. Video evidence may be used both to apprehend suspects and to trace their movements. The ease with CCTV surveillance can be deployed has led to significant volumes of video data being collected from large numbers of cameras. Some estimates suggest there are over 4 million CCTV cameras operating in the UK; others suggest this number represents a relatively conservative estimate, and more so since recent interpretations of UK legislation change how the presence of CCTV needs to be registered. The expected continuation in the growth of data collection from this source, allied to increased image resolution and incorporation of audio capture, suggests potential challenges will arise for those wishing to identify key scenes, activities and events within several thousands of hours of footage taken using multiple cameras from different angles and perspectives.

It is increasingly possible to identify specific sets of information directly from video frames, and to use such information for queries against video collections. Pure video processing to date has achieved some specific successes in recognising objects in very specific scenes, though bridging much of what researchers refer to as the Semantic Gap - understanding the relationships between identified objects and things in the real world - is still required. Such consideration of bridging has been a cornerstone of the DARPA-sponsored TRECVID initiative (Smeaton, Over and Kraaij 2006). It has also led to considerations of automatic video annotation using extant textual descriptions. Analysis of these descriptions is used for training annotation systems to recognise objects and events within unseen video footage. The derivation of key concepts and their terms from these texts and other related, or collateral (Srihari 1995), texts is used in combination with information extracted from the visual scenes. Current research into video annotation, and particularly this kind of auto-annotation, is largely undertaken within the rubric of so-called "multimedia ontologies". A brief consideration of some of the work in this area is presented in Table 1, and a further review of such work can be found in Hare et. al (2006) that covers, for example, Mori et. al, (1999) using a co-occurrence model for keywords and low-level features of rectangular image regions, approaches that segment images into regions and so on.

In REVEAL, we are using techniques for terminology extraction to highlight keywords in expert transcripts of video footage and populate a CCTV ontology for potential use in automatic metadata extraction. The system uses the commonly available GATE software. The description of the framework for linking the ontology with Computer Vision technology has been described previously (Vrusias et. al, 2007). Similar approaches include that of Jaimes and Smith (2003) who process text, available with multimedia, and extract terms using combinations of word frequency, TFDIF and entropy, and application of stemming. Resulting keywords are used in the manual construction of an ontology. The relationships between elements of the ontology are found using an algorithm for discovering association rules, and relevant relationships are manually selected and incorporated. Jaimes and Smith claim that constructing ontologies using purely textual information is not adequate: it is unclear where they consider the inadequacy to lie, though they note that issues such as different correct specifications for the same domain may be a factor. Elsewhere, free text descriptions, such as image captions, are considered as valuable sources of annotation information, not least in applications such as Google Images.

In this paper, we describe the ontology extraction components of the system in development and use in REVEAL for automatically populating a CCTV ontology using analysis of expert transcripts of the video footage. In Section 2, we describe this system; Section 3 outlines some analysis undertaken on parallel annotations; Section 4 considers future work related to the project.

	Soccer Ontology	**TRECVID**	**REVEAL**
Domain	Football	Broadcast News	CCTV
Analysis	Video	Video	Video + Text
Algorithms employed	Fuzzy c-means clustering algorithm, sum of all Needleman-Wunch distances, Bertini's annotation algorithm	Gabor texture (image properties), Grid Color Movement (colour distribution), Edge Direction Histogram (geometric cues)	**Video:** Motion detection, motion tracking, colour classification, geometric classification and object classification. **Text:** Linguistic, Statistical, Synonym and Known Terminology Analysis
Resources used	Videos of 3 football matches (World Cup & UEFA), MPEG-7, OWL	Kodak Video Collection (manually annotated), YouTube videos (manually annotated), MPEG-7, OWL	CCTV footage, transcripts of police descriptions, MPEG-7, OWL
Publication	Bertini et. al, 2007	Chang et. al, 2007	Vrusias et. al, 2007

Table 1: Recent work on multimedia ontologies involving videos and text.

2. Ontology Extraction Subsystem

The multimedia ontology in REVEAL is intended for documenting the semantics of video scenes that cannot easily be captured through video analysis techniques alone. REVEAL uses text analysis to learn this ontology from the expert descriptions, and associates elements of the ontology through the WordNet lexical database and through the Police Information Technology Organisation (PITO), now the National Policing Improvement Agency (NPIA), terminology.

Here, we describe components devised and used for REVEAL, and integrated with GATE, for the purposes of ontology learning. These components build on existing GATE plug-ins from ANNIE, for preliminary NLP tasks of such as POS tagging and sentence splitting. We use OWL for representing the ontology which captures the weight of the co-occurrence between objects and actions. The weight is calculated using frequency and distance between terms to determine the strongest relationships between objects and actions. This information is used to associate video objects with concepts from the ontology for potential use in video annotation and keyword related search expansion.

The pipeline for these resources is shown in Figure 1, with brief descriptions of each component in the remainder of this section as follows:.

- Linguistic Concept Identification (2.1)
- Statistical Concept Identification (2.2)
- WordNet Verification (2.3)
- PITO Terminology Analysis (2.4)
- Event Mapping (2.5)

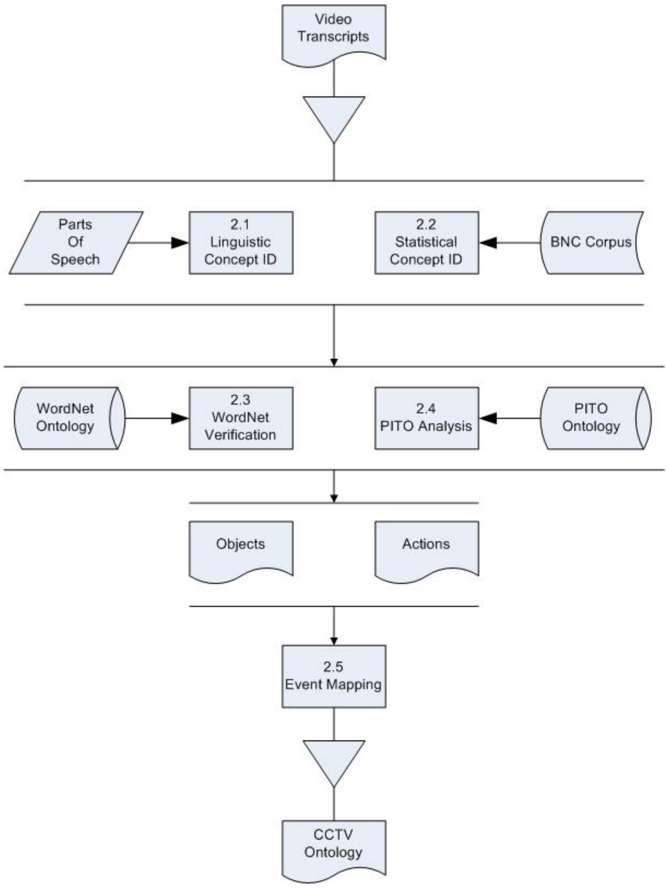

Figure 1: Pipeline for the automatic generation of the term-based CCTV ontology

2.1 Linguistic Concept Identification

We use the ANNIE tokeniser and Part of Speech tagger to identify all the nouns and verbs in the transcripts, and to provide a basis for the identification of compound nouns according to specified patterns of part of speech annotations (extended from e.g. Jacquemin 2001, p27). 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.

2.2 Statistical Concept Identification

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 Part of Speech tagger. Transcripts undergo statistical analysis to determine salience using frequency and weirdness information, as outlined by Gillam (2004), to provide statistical evidence for concepts and to act as confirmation for the linguistic extraction. Results are classified as either objects or actions via WordNet Verification (2.3). The British National Corpus (BNC) is used as a reference corpus, and thresholds are adjusted by modifying parameters for the distributions. To provide for inspection of these data, we visualize frequency and weirdness values using two-dimensions for Tag Clouds: colour and size represent weirdness and frequency respectively. Large text and red font indicates high frequency and high weirdness; small text and blue font to shows lower frequency and lower weirdness. This technique provides an overall sense of the content, and helps to identify issues such as character handling, problematic tokenization, and otherwise "unclean" data, as demonstrated in figure 2.

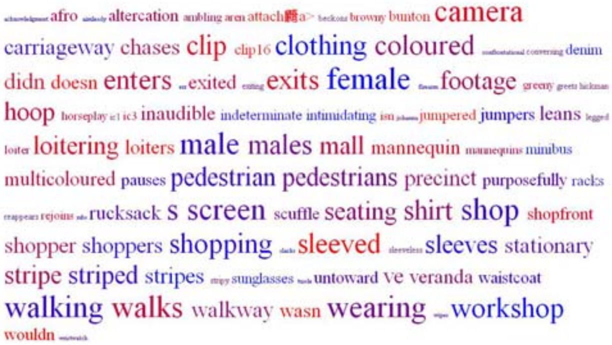

Figure 2: Tag cloud presenting the top 100 'weird' words and their frequencies in video transcripts

2.3 WordNet Verification

We use Wordnet via the Java JWordNet API to assess results of PoS tagging and compound identification (2.1). Each word tagged as a noun is checked in the WordNet noun dictionary. For initial assessment, nouns not found are considered to be erroneously identified and are rejected. Compound nouns are also verified using WordNet: if the compound noun is not found in WordNet, the first word is removed and the compound noun re-tested. For example, 'first storey shopping centre' is not known to WordNet, so 'storey shopping centre' is tested. This phrase is also not known, so we test 'shopping centre', which is present in WordNet. The process currently removes from left to find the root of the compound noun. Using WordNet this way ensures that the longest multiword expressions that represent semantic units are used for object identification. Each word tagged as a verb is also checked and rejected if it is not present. Those that are verified are used to populate the ontology as instances of Object and Action classes respectively. Concepts identified through statistical analysis (2.2) are similarly used.

2.4 PITO Terminology Analysis

REVEAL uses information and procedures defined by the National Policing Improvement Agency (NPIA), formerly known as the Police Information Technology Organisation (PITO). 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. For example, 'car' including subtypes 'saloon' and 'estate' are present in this ontology: while saloon is present in WordNet, estate is not. This information is used to associate keywords for more efficient relationship analysis in Event Mapping (2.5).

2.5 Event Mapping

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.

3. Results

To demonstrate the approach, we consider 6 expert commentaries on 12 video clips between 60 and 180 seconds showing traffic and pedestrians moving around a single lane carriageway usually with a central reservation. An image from one of the video clips can be seen in Figure 3 and two examples from the commentaries in reference to the scene are shown in Figure 4.

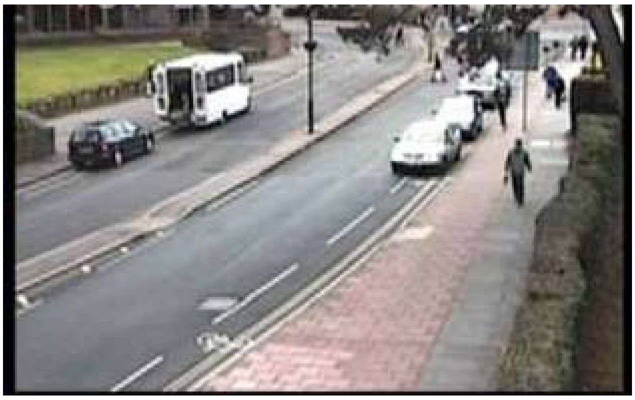

Figure 3: Scene from video

"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 4: Two examples from six transcripts describing the same scene as featured in figure 3

The transcripts demonstrated limited agreement between the experts: the same object was described in numerous different ways. A 'white van', 'kind of van', 'stationary vehicle' and 'ambulance' were all used to describe the same object in the scene. Despite inconsistencies in the descriptions, there were some frequent objects and actions identified through application of linguistic (2.1) and statistical (2.2) processing. The most frequent objects and actions in the transcripts are detailed in Table 2.

Rank	**Object**	**Count**	**Action**	**Count**
1	van	410	park	161
2	road	396	come	137
3	car	350	have	133
4	side	310	walk	126
5	pedestrian	188	cross	105
6	vehicle	149	pull	93
7	people	96	go	87
8	street	61	see	87
9	camera	51	be	70
10	person	48	get	59
Totals		**2059**		**1058**

Table 2: The top 10 identified objects and actions from the transcripts

Frequent object and action, 'car' and 'park', co-occur 65 times, with twelve variations. We can contrast these occurrences with the British National Corpus[1] to try to determine any significance. We are dealing with relatively small numbers (17098 tokens) in the transcripts, but still find use of patterns that are infrequent in 100,106,029 tokens of the BNC ('car is still parked'; 'car which was parked'; 'car that is parked') as shown in Table 3. It should be noted that it is the associations derived that are of interest.

Collocation	**Transcript Count**	**BNC Count**	**Indicative Weirdness**
Parked car(s)	39	131	2.28E-03
Car(s) parked	7	130	4.09E-04
Car that is parked	4	0	2.34E-04
Car(s) is/are parked	3	15	1.75E-04
Car(s) parking	3	220	1.75E-04
Car(s) parks	2	205	1.17E-04
Car is still parked	2	1	1.17E-04
Car which was parked	1	1	5.85E-05
Car(s) that was parked	1	1	5.85E-05

Table 3: Statistical analysis of the most frequent variations of the 'car' and 'park' collocation

The relationships between concepts determined through event mapping (2.5) are detailed in Tables 4 and 5, showing the AO Events and the AAR Events respectively. The 'Key' refers to the ranks of object and action seen in Table 2; the inter-annotator agreement column shows the percentage of expert transcripts featuring the relationship, though it doesn't imply that annotators would necessarily disagree about using these descriptions. Similar collocations can be used to strengthen assumptions regarding synonymy and semantic relationships, particularly with 'park-car' and 'park-vehicle' describing the same scene.

Key	**AO Event**	**Count**	**Inter-Annotator Agreement %**
1-3	*Park-Car*	*65*	*100*
5-2	Cross-Road	63	100
2-3	Come-Car	47	83
4-5	Walk-Pedestrian	45	83
1-1	Park-Van	35	100
2-2	Come-Road	30	83
6-3	Pull-Car	28	83
7-2	Go-Road	27	83
1-6	*Park-Vehicle*	*26*	*50*
7-3	Go-Car	23	67

Table 4: The top 10 identified AO events from the transcripts

[1] http://view.byu.edu

Key	AAR Event	Count	Inter-Annotator Agreement %
3-2-2	Car-Come-Road	20	67
10-5-2	Person-Cross-Road	16	67
5-5-2	Pedestrian-Cross-Road	11	50
3-7-2	Car-Go-Road	9	33
7-5-2	People-Cross-Road	9	50
7-3-38	People-Have-Discussion	7	17
5-4-4	Pedestrian-Walk-Side	6	50
3-2-4	Car-Come-Side	5	33
7-4-8	People-Walk-Street	4	50
19-5-8	Somebody-Cross-Street	4	17

Table 5: The top 10 identified AAR events from the transcripts

An expert referring to a ‘car’ as a ‘saloon’ leads to two separate AO events: ‘Park-Car’ with a frequency of 65 and ‘Park-Saloon’ with a frequency of 2. Since PITO Analysis (2.4) and Wordnet both identify ‘saloon’ as a type of car, we have confidence in establishing this association within REVEAL, and leading to improved inter-annotator agreement from Event Mapping (2.5). WordNet can be used, further, to associate synonymous AAR Events ‘Person-Cross-Road’ and ‘Somebody-Cross-Street’; similarly for ‘People-Walk-Street’ and via semantic distance, ‘Pedestrian-Cross-Road; ‘People-Get-Car’ and ‘Driver-Get-Vehicle’.

Erroneous relationships exist such as ‘Pedestrian-Cross-Van’ derived from the sentence, ‘Pedestrian crossing between the van and the other car’. These relationships are filtered out as frequency of events increases through analysis of further transcripts.

Further work needs to be done to investigate, amongst other things, use of active and passive voice. Currently distinct AAR events will be produced for the active and passive voice for ‘the driver parks the car’ and ‘the car is parked by the driver’.

4. Conclusion

The ontology indicates the semantics being used in video descriptions that could be used for automatic annotations, and is intended to assist in semantic transcoding of a video scene. Domain experts tend to use a language with characteristics of a specialist language or sublanguage, when describing video scenes. Words of this language can be combined with the PITO/NPIA ontology and Wordnet to provide a controlled language for video annotation. The identification of frequently used objects and actions can provide a basis for additions into the PITO/NPIA terminology for assisting in video commentary and annotation. Work in the area of terminology enhancement has already been undertaken in relation to controlled authoring (Gillam and Newbold, 2007).

Research is being undertaken to explore uncertainty issues in the semantic web (Stoilos et, al, 2006). Some applications in multimedia processing, such as object recognition, need to process incomplete or random information. It has been suggested that these applications use some form of probability measurement, with a fuzzy extension to OWL being proposed. The event mapping techniques described here with frequency and inter-annotator agreement ratings could provide a basis for such a probability measurement. In addition, Del Bimbo (1999) described how new-generation video retrieval systems are commonly accessed by users through browsing and navigation and how the relevance feedback on search results can be used to refine annotations for future search queries. This feedback can be used as additional weighting for such a probability measurement. The resulting process could demonstrate aspects of the human cognitive ability to associate visual experiences with semantic information.

The ontology can be extended by assigning multimedia objects to concepts. Bertini et. al (2007) showed how multimedia ontologies can be used to perform automatic annotation on unknown sequences of video. Using visual descriptors and temporal cues, a representative set of sequences containing concepts described in the linguistic ontology can be used to create a multimedia ontology. By checking the similarity of visual descriptors of unseen video with the representative sequences, automatic video annotation can be achieved. Figure 5 shows the relationships to events for scenes using a multimedia ontology defined for CCTV. This is a starting point for automatically creating annotations for video and other forms of multimedia information in the semantic web, and this work in REVEAL is ongoing.

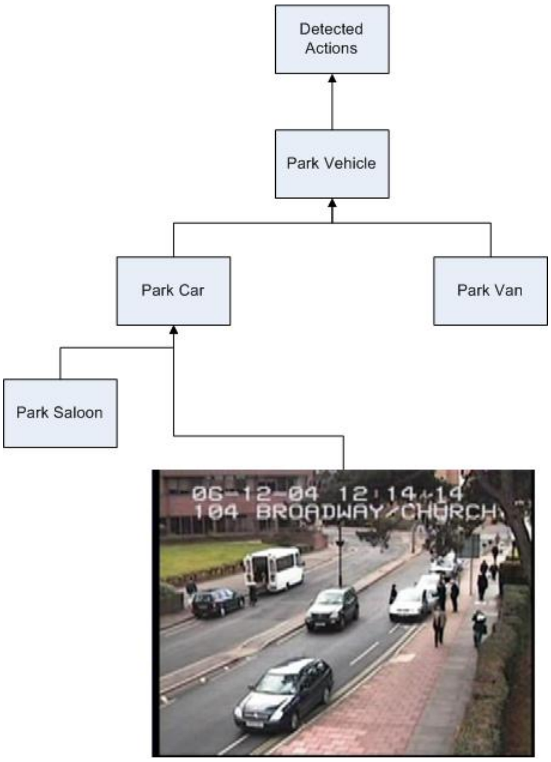

Figure 5: Schema used to annotate CCTV video clip

5. Acknowledgements

This work has been supported, in part, by EPSRC-sponsored REVEAL project (GR/S98450/01) and by the EU eContent project LIRICS (22236).

6. References

Bertini, B. Del Bimbo, A. D. and Torniai, C. (2007). Soccer Video Annotation Using Ontologies Extended with Visual Prototypes, In *Proceedings of the International Workshop on Content-Based Multimedia Indexing, CBMI '07*, vol., no. 25-27, pp. 212--218.

Bimbo, A. D. (1999). *Visual Information Retrieval*. Morgan Kaufmann.

Chang, S., Ellis D., Jiang, W., Lee, K., Yanagawa, A., Loui, A. C. and Luo, J. (2007). Large-Scale Multimodal Semantic Concept Detection for Consumer Video, In *Proceedings of the international workshop on Workshop on multimedia information retrieval,* Augsburg, Bavaria, Germany, pp: 255--264.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics(ACL'02).* Philadelphia, July 2002.

Gillam, L. (2004). Systems of concepts and their extraction from text. Unpublished PhD thesis, University of Surrey

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

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing.* The MIT Press.

Jaimes, A. and Smith, J.R. (2003). Semi-automatic, data-driven construction of multimedia ontologies. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*. Volume 1, 6-9 July 2003, pp. 781--4.

Mori, Y., Takahashi, H., and Oka, R.. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99).*

Newbold, N. and Gillam, L. (2007). Automatic Document Quality Control. *In Proceedings of the Sixth Language Resources and Evaluation Conference (LREC).* Marrakech.

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval,* Santa Barbara, California, USA, October 26 - 27, MIR '06. ACM Press, New York, NY, pp. 321--330.

Srihari, R. K., (1995). Use of Collateral Text in Understanding Photos, *Artificial Intelligence Review*, special issue on Integrating Language and Vision, Volume 8, pp. 409--430. [* Also reprinted as book chapter in Paul McKevitt (ed), Kluwer, 1995.]

Stoilos, G., Simou, N., Stamou, G. and Kollias, S. (2006) "Uncertainty and the Semantic Web", *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 84-87, Sept/Oct, 20.

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