Visual Surveillance and Video Annotation and Description

Khurshid Ahmad, Craig Bennett & Tim Oliver

Department of Computer Science, Trinity College, Dublin-2, Ireland
Department of Computing University of Surrey, Guildford, Surrey, UK
Khurshid.Ahmad@cs.tcd.ie, c.bennett@surrey.ac.uk

Abstract

The effectiveness of CCTV surveillance networks is in part determined by their ability to perceive possible threats. Our traditional means for determining a level of threat has been to manually observe a situation through the network and take action as appropriate. The increasing scale of such surveillance networks has however made such an approach untenable, leading us look for a means by which processes may be automated. Here we investigate the language used by security experts in an attempt to look for patterns in the way in which they describe events as observed through a CCTV camera. It is suggested that natural language based descriptions of events may provide the basis for an index which may prove an important component for future automated surveillance systems.

1. Introduction

In the UK, and perhaps all across the European Union, a large number of agencies collect video images for the purpose of protecting life and property. Life and mission critical situations require a rapid retrieval of images held in the video archives, as has been demonstrated in the various emergencies in major European cities.

The collection and analysis of video images, to date, has been the preserve of computer vision experts on the one hand, with their focus on low-level visual features, and vision/experimental psychologists with their focus on high-level, image external features, including naming and description of (complex) objects – a *suspicious person*, an *abandoned bag* - in a video stream.For browsing and searching a video index it is important to have an *index*. Three questions have been asked in this context: what to index – some frames all of the frames; how to index – the application of pattern classifiers to the auditory or visual channels; and, which index – names of persons, places or things, the spatio-temporal location of an object (Snoek and Warring 2005). The authors discuss the creation of a 'semantic-index': this index comprises meta-linguistic properties of a video stream, including *genre*, *purpose*, and its *logical units* (a continuous part of video that can be named to express a particular object or event). The index has named events wherein visual features are marked with high level (named) events (ibid:21).

Increasingly, it has been argued that the videos have to be annotated using keywords for creating a semantic index. However, the keywords are arbitrary selected on a domain-by-domain basis: videos comprising suspicious and perfectly normal images have been annotated by a selection of verbs – *walking, running, browsing* (Fisher et al 2004); medical video streams are annotated by experts using their intuition, but there are reports that unified medical language system UMLS is used for medical terminology; traffic videos typically focus on number plates recognition and speeds of vehicles are the two underlined nouns and their broader and narrower versions are used in annotating traffic videos.

More ambitious projects have investigated the 'generation of natural language descriptions of human behaviour from videos' (Kojima et al 2000). The authors use Fillmore's case grammar and an assortment of vocabulary and intuitively chosen verbs: the claim here is that 'through applying word dictionaries (sic), case structure patterns of verbs, and syntax rules into a behavioural expressions, natural language text can be generated' (ibid:2000:73 1).

2. Overview

In this paper we describe our attempts for automatically indexing video streams from a set of specially created training data. The training data comprises a set of video sequences together with the commentaries of four different experts who volunteered to attend a meeting after an invitation was extended at a series of workshops on video surveillance held in 2005. Our investigation focussed on the identification and 'discovery' of idiosyncratic lexico-grammatical patterns. At the level of lexis, the inter-indexer variability was present in infrequently used verbs of motion and action, but it appears that there is a small set of frequently used verbs that is used extensively amongst the experts. At the level of grammar, the inter-indexer is perceptible when we 'parsed' the sentences in the experts' commentaries into phrases containing action and result (of an action), and location of an object in the video stream: A large number of sentences in the description of an event comprise words that can be classified as verbs of action, nouns related to location, and verbs that are results of actions (ALR), a finite state automata

L?M?((A+L)M?)+L?A*M?R,

where the superscript '+' indicates recursion ($A+ \rightarrow A$, or AA, or AAA so on), and M denotes a phrase that does not contain terms relating to ALR; only sentences – a majority of sentences in our corpus of descriptions- ending in R were selected. The automata retrieves 72±12.5 of the patterns found in the description of our four experts correctly with the 'ambiguity' M present.

3. Data and Method

In this section we describe the dataset used in our experiments and a four-step method for analysing the commentary of a video sequence.

3.1. CAVIAR Dataset

The 16 clips for which the annotations are reported, 8 are subset of the PETS 2004 CAVIAR INRIA dataset [1], with the remaining 8 clips originating from the extended CAVIAR Lisbon dataset. The INRIA dataset consists of 28 video of between 500 and 1400 frames and totalling 26500 frames, filmed in the lobby of the INRIA Labs at Grenoble, France. The Lisbon dataset was collected from a hallway in a shopping centre in Lisbon and contains 26 scenarios. On average each clip within consists of 1500 frame, and contains larger numbers of people and groups of people than the INRIA dataset.

The subject of the videos is orientated about the surveillance task, and is grouped into 12 broad categories including: walking, browsing, resting, meeting, leaving objects, and fighting. Accompanying each frame there is ground truth information annotated by hand, including such information as location for major actors and a limited linguistic description of events. The ground truth data is specified in XML and is based on the CVML language [2] and both video and ground truth data are publicly available for download2.

3.2. Method-I

Protocol Analysis of the CAVIAR Dataset: Within our experiments 4 subjects participated, all with a background in policing who were asked to comment on a collection of surveillance videos. The subjects were invited to verbally describe events as they took place and highlight any unusual incidents, their comments were subsequently transcribed. A total of 16 videos were presented to the subjects from the CAVIAR dataset which were displayed at a resolution of 384 by 288 pixels. After a period of microphone calibration the videos were shown to the individuals successively and their comments were recorded. We have used a video library of 28 sequences that are 'orientated about a public surveillance task' and cover six test scenarios – walking, browsing, collapsing, leaving objects, meeting and fighting (Fisher 1996). Protocol analysis techniques were used and six video surveillance experts, drawn from the UK police forces and government research and development organisations, described the public surveillance videos. Note that Fisher's video sequences (aka PETS2004 data set) are 'ground truth labelled' frame-by-frame with bounding boxes and have been annotated by a semantic description of the six test scenario activities mentioned above. The ground-truth serves as a base line for the analysis of the experts' descriptions.

The expert commentaries were taped and transcribed by a trained audio typist. The size of the commentary corpus is over 50,000 words.

3.3. Method-II

Pre-processing of the Commentary: It was necessary to perform pre-processing of the experts' commentaries – essentially transcriptions of free, spontaneous natural language speech output: First, the descriptions of each of the 16 videos were separated by hand such that each video commentary was rendered as a 'paragraph'. Second, there are instances where two or more separate events were depicted in the video sequences: these events were separated by hand again and treated separately.

3.4. Method-III

Action-Location-Result Markup: One can argue that a given event comprises one or more actions in a certain location that leads to a 'result' or consequence. Our assumption is that within a 'sentence' that describes an event one can mark-up phrases that relate to the action-location-result (ALR) triple; what cannot be marked up consistent with the intuitively outlined triple will be marked-up as miscellaneous (M). We show that one of the most frequent triple is 'LAR' and there are many complex structures in which the triple either is embedded or there are sequences that are embedded between one or more elements of the triples.

3.5. Method-IV

Frequency of usage a particular unit at a given level of linguistic analysis relates to the acceptability of the unit within a linguistic community: This is a dictum on which Randolph Quirk created the corpus-linguistic movement. We have analysed the distribution of single words in our commentary corpus. Each word was assigned a grammatical category either the use of the University of Lancaster's CLAWS tagger (http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/) and by looking up the OED On-line (http://www.oed.com/): typically we have used categories as verbs of action, motion and stasis and combinations thereof with adverbs+particles, prepositions and with nouns. We have used only one morphological form a given verb, e.g. {walk, walked, walking, walks} were all marked up as walk.

It is essential for the existence of a local grammar of video description that the experts use a small number of verbs frequently and preferentially. The same is true of location words with an added dimension of ambiguity in that experts usually describe 'real world' locations, for instance stairs and walkways, and use phrases to point out spatial locations relative to locations – to the left of the stairs, the area on the left-- but as they are watching a vide sequence they sometimes refer to the locations in relation to the portrayal of the video sequence – the right hand side of the screen, away from us.

The most frequently used tokens for action, location and result are used as anchors for building a regular expression.

4. Analysis

4.1. Lexical Analysis

Typically a commentary will start as :

Man in blue t-shirt, centre of scene, facing camera, raises white card high above his head.

The above excerpt was marked up as:

Event 1:	M	Man in blue t-shirt,
	L	centre of scene,
	A	facing camera,
	R	raises white card high above his head.

The lexical level analysis then focussed on each of the constituent triples (L,A, R) was carried out and then followed by an analysis of the each of the commentaries at sentential level.

Lexical Level of the description of the commentaries: The commentaries of the four experts were analysed and we found the distribution of the verbs of motion, action and stasis in the 'action' sequence of the commentaries was 61%, 29% and 7% and all other types of verb combinations accounted for just under 2% of the total. A lexical level analysis showed that 'walk' dominates the commentaries of our experts (60%) followed by 'enter' (14%) and 'stand' (4%) of all the verb tokens used (see Table 1a below)

Expert	E1	E2	E3	E4	
Walk	40	56	40	29	61%
Enter	1	6	19	13	14%
Stand	2	1	5	3	4%
Meet	1	3	1	4	3.3%
Sit	1	1	3		1.8%
Come together		1	1		0.7%
Look	1		4		1.8%
Exit	1	1			0.7%
Stop	2		1		1.1%
Total	49	69	74	49	241
Unique Words	4	8	10	8	30
Grand Total	53	77	84	57	271

Table 1a: The distribution of verb tokens used in the action sequence of the commentaries; absolute frequencies are used and the percentage is based on the total number used.

All our experts use verb tokens not used by others – but this number is small as shown in Table 1b:

E1	Interact (3) Approach (3); Stop; Raise arms; Make intimidating gestures; Make contact; Leave
E2	Struggle; Run off; Pick up; Part; Get to feet; Fight; Deposit; Check time
E3	Converse (2); Spend Time; Face; Confront; Browse; Arrive; Altercation
E4	Wait; Hit; Get into; Appear;

Table1b: The 'unique' verbs used by experts; numbers in parentheses indicate the number of time the token was used more than once.

The distribution of verbs in the 'result' sequences of the commentaries is marginally more diffuse when compared to the 'action' sequences of the commentaries: (Table 2)

Expert	E1	E2	E3	E4	
Walk	8	19	9	18	47.4%
Exit	7		3	2	3.5%
Run	2	2	2		10.5%
Look	1	1	1	1	5.3%
Follow	1	1	1		2.6%
Stand			1	1	1.8%
Get up	1		1		1.8%
Chase	1		1		1.8%
Raise			1	1	1.8%
Put		1	1		1.8%
Meet		1		1	1.8%
Retrieve		1	2		2.6%
Total	21	26	23	24	94
Unique Words	2	1	4	13	20
Grand Total	23	27	27	37	114

Table 2: The distribution of verb tokens used in the 'result' sequence of the commentaries; absolute frequencies are used and the percentage is based on the total number used.

The number tokens used that are unique to a given expert are under 20% of all verb tokens used here –the figure was 11% in the action sequences. The distribution of location words showed a dominance of reference to 'real world' objects (55%) with the balance between locations relative to portrayal (24%) and to relative spatial location (20%).

4.2. Sentential Level Analysis

Consider the analysis of one of our videos and the first five sequences in the video stream: Most descriptions begin with an L or A, whereas few begin with M and none begin with R. It is also apparent that L's and A's tend to appear together, with many sequences of AL and LA but few AA or AL. This initial analysis provides motivation to further analyse the data in order to quantify any common patterns that exist.

1	L	A	L	A	A	R	M	A	R	A
2	A	L	A	L	A	L	R	L		
3	A	L	M	R	L					
4	A	L	A	L	R	A				
5	A	L	A	R	L					

Table 3: Sentence structure in terms of markup tokens

A look for larger distribution patterns indicates that there is a dominance of certain patterns over the others as an averaged distribution of triples shows below:

ALA	LAL	ALM	ALR	MAL
30.75	19	14.25	9.75	8

Table 4: Pattern distribution of token combinations

The most popular triplet that ends with an R is ALR, therefore the data containing this triplet was selected for further analysis. Discarding the rest of the data and sorting on the columns to the left of the ALR shows further similarity between all of the experts. The most common sequence was ALALR, but ALALALR also existed, as did ALALALALR. This suggests that the grammar permits the repeated use of AL, so long as it is

followed by an R. There were also several instances where more than one A appeared together, for example AALALR amongst others, but at no point was L used more than once consecutively. The following regular expression was constructed from our analysis that allows for recursive patters (AAA…ALR, ALALALR and so forth):

L?M?((A+L)M?)+L?A*M?R,

5. Results

The regular expression was then used to extract the sequences in our corpus. Note that the regular expression –our local grammar- indicates that all sequences should end in the result (R). There are a number of sequences that do not and these are not retrieved by our local grammar. However when we ignore all patterns that do not end in the result then our overall results improve significantly:

Expert	Accuracy: All sequences (%)	Accuracy: Sequences ending in 'R' (%)
1	86	100
2	79	93
3	63	77
4	60	67

Table 5: Accuracy of our approach for all sequence and sequences not ending in token 'R'

The results of the distribution at the lexical level and at the regular expression (or sentential level) indicate that Experts 1 & 2 are quite close whereas Experts 3&4 use larger number of tokens (both in action and result sequences, see Tables 1a and 2). Consequently, our regular expression does not capture the sequences Experts 3&4 describe as well. We are continuing with our analysis of the two other experts and attempting to automate the 'discovery' of the various verbs and nouns. We are encouraged by our results.

Acknowledgements

This work has been supported by the UK Engineering and Physical Sciences Council's grant REVEAL, (GR/S98443/01). The project is being conducted in close collaboration with Kingston University and Sira Ltd, and supported by Police Information Technology Organisation (PITO) and Police Scientific Development Branch (PSDB). The work reported here was carried out by Tim Oliver for his MSc dissertation supervised by Khurshid Ahmad; Craig Bennett helped with generation of experimental data.

The CAVIAR project is EC funded project/IST 2001 37540 conducted at the University of Edinburgh.

6. References

CAVIAR (2005)
Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ Last accessed: 12th October 2005

Fisher, R. (2004) PETS04 Surveillance Ground Truth Data Set, *Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance(PETS04)*, 1-5

Kojima, M. Izumi, T. Tamura and K. Fukunaga (2000) Generating Natural Language Description of Human Behaviour from Video Footage, *In Proc. 15th Int IEEE Conference on Pattern Recognition (ICPR'00)*, 4, 728–731

Snoek, C., Worring, M. (2005) Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*