Visual Semantics for Reducing False Positives in Video Search

Rohini K. Srihari

Dept. of Computer Science & Engineering
State University of New York at Buffalo

rohini@cedar.buffalo.edu

Adrian Novischi

Janya Inc.
Amherst, NY

anovischi@janyainc.com

Abstract

This research explores the interaction of textual and visual information in video indexing and searching. Much of the recent work has focused on machine learning techniques that learn from both text and image/video features, e.g. the text surrounding a photograph on a web page. This is useful in similarity search (i.e. searching by example), but has drawbacks when more semantic search is desired, e.g. find video clips of Obama meeting with ordinary citizens. By extracting key visual semantics from the audio/text accompanying video, we are able to enhance the precision and granularity of video search. Visual semantics relates to identifying and correlating linguistic triggers with visual properties of accompanying video/images. Significant progress has been made in text-based information extraction, which can be brought to bear for video search. In this paper, we focus on linguistic triggers related to a special class of events referred to as nominal events. We describe how proper detection and interpretation of such events can prevent false positives in video searches.

1. Introduction

This paper investigates the interaction of textual and photographic information in a video and image search system. With the rising popularity of websites such as YouTube, the need for semantic and granular search of video content is apparent. Currently such material is being indexed purely by metadata and keywords provided by the users contributing the content. While this is often sufficient for ranking video clips, it is not sufficient if more granular search of video content is necessary, i.e., searching for specific segments within a long clip corresponding to a specific event. Even in such cases, there is often text/audio content accompanying the video, (e.g. voiceovers) that can be exploited for more semantic indexing and search. Applications such as question-answering have focused on granular text search, whereby the unit of retrieval is a phrase or a text snippet; video search is beginning to require similar levels of granularity, especially when mobile, small-form devices such as the iPhone are being used.

There has been considerable research in the area of multimedia indexing and retrieval. Much of the early focus was on content-based indexing and retrieval of images (Smeulders et al., 2000). This body of work was concerned with using image features such as colour, texture, shape, etc. in indexing images. Evaluation was typically on a large collection of images such as the Corel data set; queries were typically similarity-based, where an image was used as a query. It is felt that most of the achievable gains in image-based retrieval have already been realized. The new frontier focuses on exploiting the temporal dimension in content-based video indexing and retrieval (CBVIR) in order to detect entities and events; several DARPA and NIST programs have focused on this task (Smeaton et al., 2006). This task is extremely challenging since, in most cases, no collateral audio/text is available. PICTION (Srihari and Burhans, 1994) was a system that identified human faces in newspaper photographs based on information contained in the associated caption. It was noteworthy since it was (i) one of the first systems to combine both text and image data in semantic image search, and (ii) provided a computational framework for text-guided image interpretation in situations where pictures are accompanied by descriptive text. More recent work (Feng and Lapata, 2008) focuses on the use of machine learning techniques that combine both text features (such as bag-of-words) along with video/image features in joint indexing of video/images. The assumption is that this will: (i) allow for both text-based and similarity-based search, and (ii) eliminate the need for manual video/image annotation. While the initial results have been impressive, they have not focused on the problem of granular, semantic video search. Text features have typically been restricted to variants of bag-of-words features and have not exploited advances in information extraction technology. Thus, a search for video clips of Clinton giving a speech will also result in video clips of an empty hall; the latter would have been retrieved due to the accompanying audio saying volunteers are getting the museum building ready in preparation for Clinton's speech tonight.

The focus of this paper is on exploiting advances in text-based information extraction for more granular and accurate searching of video. In particular, we focus on reducing false positives such as the above by finding key linguistic triggers in the form of nominal events. We first

discuss some initial results in detecting events in video. The techniques are based on a traditional text-based IE system. We revisit the theory of visual semantics (Srihari and Burhans, 1994), a framework used in situations where visual and textual/audio material are jointly presented to a user: visual semantics is concerned with correlating linguistic structures with the semantics of a scene. Semantics of a scene include the objects present, the events taking place, background context (e.g. scenic background versus city setting). The focus of this paper is on and how they contribute to accurate searching of video/images by filtering out noise. Such a text-based technique can be eventually combined with the advances made in CBIVR to enable the most accurate video search possible.

2. Event Detection in Video

This section describes how a text-based information extraction system was customized to find events of interest in video. The text transcripts of audio accompanying video were indexed by the Semantex (Srihari et al., 2006) engine. Semantex is a hybrid engine for information extraction, incorporating both machine-learning and grammatical paradigms. It is capable of extracting entities, events, and relationships and attributes, as well as other nontraditional IE output such as sentiment analysis. Events of interest included marriage, divorce and stealing. Events were extracted from a corpus of about 300 documents containing about 872 KB of text. In this corpus, Semantex automatically found 191 instances of marriage events, 59 instances of divorce events and 94 instances of stealing events. The accuracy of event detection for each type is presented in ***Table 1***.

Event Type	Correct	Attempted	Accuracy(%)
Marriage	178	191	93.2%
Divorce	56	59	94.9%
Stealing	77	94	81.9%

Table 1: Accuracy of Semantex system in finding three types of events: marriage, divorce and stealing

We found that some of the detected events that are mentioned in the audio transcripts do not correlate with events presented in the video. A simple indexing based on keywords would lead to retrieval of false positives. For example in the following paragraph:

> *"I, unfortunately, in the last couple years, did not have any conversations with her because she was withdrawn so much," said Becky Best, who was maid of honor at the couple's 1990 wedding. The couple had met through a dating service a few years before their marriage, she said.*

The system found the phrase *at the couple's 1990 wedding* as being an instance of a marriage event. Although this is counted as a correct detection of a wedding event in text, it unfortunately does not correspond to a wedding event in an accompanying video clip. Later sections discuss this discrepancy and how it can be avoided.

3. Visual Semantics

Visual information in collateral text answers the questions who or what is present in the accompanying scene and provides valuable information on how to locate and identify these objects or people. When combined with a priori knowledge about the appearance of objects and the composition of typical scenes, visual information conveys the semantics of the associated scene. The resulting semantics provides the basis for semantic searching of video and images. We refer to the theory of extracting visual information (and the associated semantics) from text as visual semantics.

3.1 Entity Classification

A primary task is to correctly detect and determine the semantic category of key entities in the accompanying video. Examples of such classes include person, place, organization, etc.. Considerable progress has been made on accurate tagging of named entities; considerable work remains in tagging and categorizing nominal entities such as the protestors, the herd, etc. As an example of how difficult this problem can get, consider the caption "*Winning Colors with her trainer Tom Smith prior to the start of the Kentucky Derby*" which accompanied a clip of a horse and her trainer. Determining that 'Winning Colors' is not a human and is actually a horse involves understanding the meaning of the word 'trainer' as well as contextual knowledge about the Kentucky Derby.

3.2 Linguistic Triggers useful in predicting video content

The description of a video clip may contain the mentions of several events. Most of the time only one event in its description is presented in the video clip. For example, in the video clip where *Tom Smith and his wife Mary Jane are preparing for the visit of President Clinton on Tuesday*, the visit of President Clinton is not expected to be in the video clip. Most of the concern is finding the main events and their entities and rejecting those events and entities that are not expected to appear; not all entities mentioned in the accompanying description are in the video. Predicting the event(s) described in the video together with the relevant entities can benefit from using linguistic triggers to detect relations between events. We identified three situations:

(i) **One event is the argument of the other**: This situation is possible if, by the use of reification of one event, it is expressed as a verb nominalization and becomes an argument of the other event. This is why it is important to address the detection of nominal events. For example in the following video description from the CNN website *CNN political analyst Mark Preston talks about John McCain's accusations of Barack Obama's "financing flip-flop*", the "flip-flop" event is an argument of the accusation event which in turn is an argument of the discussion event. In these cases, prepositions like "about", "of", "for" represent very good linguistic triggers for indicating that one event is an argument of another.

(ii) **Linguistic Modality**: Some events described in the text are real and others are unreal or uncertain. These events that are unreal or uncertain are introduced by modal verbs like "may", "might", "can", "could" or modal expressions like "it is possible", "it is probable", or intentional adjectives for nominal events like "alleged", "fake", "possible" "probable" (Palmer, 2001). All these are important linguistic clues to detect events and entities that are not likely to appear in the video content. As an example, in the following description from CNN website: *Possible floods force earthquake refugees in China to put what little they have onto bikes and evacuate their tents.* The "floods" event is not real and it will not be shown in the video because it is modified by the adjective "possible" (Peters and Peters, 2000).

(iii) **Temporal relations between events**: In some cases there is a temporal relation between events where the time frame of the main event in the description is given using some temporal relationship to other event. For example in the following video description from CNN website *Italian soccer fans set fireworks off after a victory against France in Zurich in the Euro 2008 games*, the "victory" event prefixed by the preposition "after" is in a temporal relationship with the main event, "setting fireworks off". Temporal relationships between events are expressed by prepositions like "before", "after" which are important linguistic triggers to select the most important event and entities.

4. Experimentation with Nominal Event Detection

Nominal event detection plays an important role in avoiding false positives in video search. This section presents our approach for detecting nominal events, inspired by the word of (Creswell et al., 2006). This approach is based on a probabilistic model called multinomial model. This model uses vectors of features computed for each word in a certain class: event or non-event. In two different instances a word can be classified both as an event and non-event, but for a given classification of a word we are interested in those features that are correlated with that classification. For a given word in a set corresponding to a certain classification, the frequencies of all the features in its vector are computed from a corpus. The multinomial model has two requirements:

1. Future instances of words with features similar to the words with a certain class, should receive the same class label (words with similar features to words that are events should be labeled as events).
2. All the words in the set of words corresponding to a certain classification should have a contribution proportional to their frequency in the training corpus.

These requirements are incorporated naturally into mixture model formalism where there are as many components as the number of words with a given label. Let Λ be the set of all the words that were assigned label L in the training corpus. The i^{th} component s^i build around word i from the set Λ is a multinomial probability of its vector of features v:

$$p(v|s^i)=\prod_{f=1}^{F}(\hat{s}_f^i)^{vf}$$

Where $\hat{s}_f^i$ is the proportion of times the word i was seen with feature f in the training corpus compared to the number of times the word *i* was seen with any feature. $\hat{s}_f^i$ is simply the (i,f) entry in row-sum normalized matrix:

$$\hat{s}_f^i=\frac{s_f^i}{\sum_{F'=1}^{F}s_{f,}^i}$$

vf represents the number of times the word i is seen with the feature f in the test corpus.

The second requirement is realized by forming a weighted mixture of the above multinomial distribution for all the words i from the set Λ. The weighting of the i^{th} component $p(s^i)$ is fixed to the ration between the number of times the word i occurred in the training corpus with label L and the total number of all occurrences of words with label L:

$$p(s^i)=\frac{|s^i|}{\sum_{i'\epsilon\Lambda}^{F}|s^{i'}|}$$

The generative probability for the vector v of features for the words in set Λ with label L is the following:

$$p(v|L)=\sum_{i\epsilon\Lambda}p(v|s^i)\ p(s^i)$$

The new instance in the test corpus, the probability for $L=EVENT$ is computed together with the probability for

L=NON-EVENT. Then the difference between the logarithm if the two probabilities (log-odds ratio) is computed:

$d(v)=\log(p(v|\text{EVENT}))-\log(p(v|\text{NON-EVENT}))$

The sign of $d(v)$ gives the label of the new instance.

The training corpus was taken to be the Foreign Broadcast Information Service. A set of 95 event seeds and 295 non-event seeds was used to automatically tag instances in this corpus as positive and negative examples. The multinomial model used as features the set of strings created from the relations $(*;R;w_t)$ output by the Semantex system. In addition we used a set of 50 topical labels computed off-line from the same corpus using Gibbs Sampler HMM-LDA model (Griffiths et al., 2005) (with the following parameters: number of syntactic states NS=12, number of iterations N=200,ALPHA=1,BETA=0.01,GAMMA=0.1). For comparison with the previous work we used the same test corpus described in (Creswell et al., 2006) consisting of 77 documents with 1276 positive instances and 8105 negative instances for a total of 9381 labeled instances. For each instance in the test corpus, we considered three ways to compute the feature values:

(i) word - the feature values are computed from the vector of features from the training corpus.

(ii) context - the feature values are computed from the current context of the word in the test set, and

(iii) word+context - for each instance we multiply the probability information from the word vector with the probability information from the context vector.

Table 2 presents the original results reported in (Creswell et al., 2006) for comparison.

Table 3 presents the results of our model for each type of test: word, context and word+context on the original data set. The multinomial model was also tested on the corpus of video transcripts. The 300 files in the corpus were split randomly into 250 files used for training and 50 files used for testing. All the 9348 nouns in these 50 files were manually labeled with an EVENT or NON-EVENT label. This annotation effort resulted in 1035 instances labeled as EVENT and 8313 instances labeled as NON-EVENT. ***Table 4*** presents the evaluation of the multinomial model of the corpus of video transcripts. The overall accuracy of nominal event detection is above 70% in the best case. These results are promising and will allow future use of lexical triggers in identifying the main events occurring in a video and thereby avoid false positives.

5. Summary

We have focused on the problem of granular and semantic searching of video, e.g. searching for specific events involving specific entities. We are interested in situations where collateral audio or text is available, thereby permitting recent advances in text based information extraction to be exploited. The problem of detecting nominal events and their significance in accurate event search has been explored in depth. Initial results are promising; although the problem is challenging, the impact on video search is expected to be significant. Our next step will be to evaluate this on a database of video clips and compare the performance to keyword-based search.

Feature Vector	EVENT			NON-EVENT			TOTAL			Average accuracy(%)
	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	
word	1236	1408	87.7%	4217	6973	60.5%	5453	8381	74.8%	74.1%.
context	627	1408	44.5%	2735	6973	39.2%	3362	8381	37.4%	41.9%
word+context	1251	1408	88.8%	4226	6973	60.6%	5477	8381	75.0%	74.7%

Table 2: Results of the original experiments reported in (Creswell et al., 2006)

Feature Vector	EVENT			NON-EVENT			TOTAL			Average accuracy(%)
	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	
word	1127	1276	88.3%	5892	8105	72.7%	7019	9381	74.8%	78.6%.
context	608	1276	47.7%	2897	8105	35.7%	3505	9381	37.4%	41.9%
word+context	1147	1276	89.9%	5903	8105	72.8%	7040	9381	75.0%	79.2%

Table 3: Results of the experiments with multinomial model using three ways to compute the feature vector: i) word, ii) context, iii) word+context

Feature Vector	EVENT			NON-EVENT			TOTAL			Average accuracy(%)
	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	Correct	Attempted	Accuracy(%)	
word	570	1035	55.1%	5916	8313	71.2%	6486	9381	69.4%	65.2%.
context	737	1035	71.2%	5075	8313	61.0%	5812	9381	62.2%	64.8%
word+context	622	1035	62.2%	6387	8313	76.8%	7009	9381	75.0%	71.3%

Table 4: Results of the experiments with multinomial model on the corpus of video transcripts.

References

Creswell, C., and Beal, M. J., and Chen, J. and Cornell, T. L. and Nilson, L. and Srihari, R. K. 2006. Automatically extracting nominal mentions of events with a bootstrapped probabilistic classifier. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 168–175, Morristown, NJ, USA. Association for Computational Linguistics.

Feng, Y., and Lapata, M. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio, June. Association for Computational Linguistics.

Griffiths, T. L., and Steyvers, M., and Blei, D. M. and Tenenbaum, J. B. 2005. Integrating topics and syntax. In *Lawrence K. Saul, Yair Weiss, and L´eon Bottou, editors, Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, Cambridge, MA.

Palmer, F. R. 2001. Mood and Modality (2nd ed.). *Cambdidge: Cambridge University Press*. I. Peters andW. Peters. 2000. The treatment of adjectives in simple: Theoretical observations. In *Proceedings of LREC*. European Language Resources Association.

Smeaton, A. F., and Over, P. and Kraaij., W. 2006. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA. ACM Press.

Smeulders, A. W., and Worring, M. M., and Santini, S., and Gupta, A. and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380.

Srihari, R. K., and Burhans, D. T. 1994. Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *Proceedings of AAAI-94*, pages 793–798. Seattle, WA.

Srihari, R. K. and Li, W. and Cornell, T., and Niu, C. 2006. Infoxtract: A customizable intermediate level information extraction engine. *Natural Language Engineering*, 12(4):1–37.