Extracting Salient Keywords from Instructional Videos Using Joint Text, Audio and Visual Cues

Youngja Park and Ying Li
IBM T.J. Watson Research Center
Hawthorne, NY 10532
{young_park, yingli}@us.ibm.com

Abstract

This paper presents a multi-modal feature-based system for extracting salient keywords from transcripts of instructional videos. Specifically, we propose to extract domain-specific keywords for videos by integrating various cues from linguistic and statistical knowledge, as well as derived sound classes and characteristic visual content types. The acquisition of such salient keywords will facilitate video indexing and browsing, and significantly improve the quality of current video search engines. Experiments on four government instructional videos show that 82% of the salient keywords appear in the top 50% of the highly ranked keywords. In addition, the audiovisual cues improve precision and recall by 1.1% and 1.5% respectively.

1 Introduction

With recent advances in multimedia technology, the number of videos that are available to both general public and particular individuals or organizations is growing rapidly. This consequently creates a high demand for efficient video searching and categorization as evidenced by the emergence of various offerings for web video searching. [1]

While videos contain a rich source of audiovisual information, text-based video search is still among the most effective and widely used approaches. However, the quality of such text-based video search engines still lags behind the quality of those that search textual information like web pages. This is due to the extreme difficulty of tagging domain-specific keywords to videos. How to effectively extract domain-specific or salient keywords from video transcripts has thus become a critical and challenging issue for both the video indexing and searching communities.

Recently, with the advances in speech recognition and natural language processing technologies, systems are being developed to automatically extract keywords from video transcripts which are either transcribed from speech or obtained from closed captions. Most of these systems, however, simply treat all words equally or directly "transplant" keyword extraction techniques developed for pure text documents to the video domain without taking specific characteristics of videos into account (M. Smith and T. Kanade, 1997).

In the traditional information retrieval (IR) field, most existing methods for selecting salient keywords rely primarily on word frequency or other statistical information obtained from a collection of documents (Salton and McGill, 1983; Salton and Buckley, 1988). These techniques, however, do not work well for videos for two reasons: 1) most video transcripts are very short, as compared to a typical text collection; and 2) it is impractical to assume that there is a large video collection on a specific topic, due to the video production costs. As a result, many keywords extracted from videos using traditional IR techniques are not really content-specific, and consequently, the video search results that are returned based on these keywords are generally unsatisfactory.

In this paper, we propose a system for extracting salient or domain-specific keywords from instructional videos by exploiting joint audio, visual, and text cues. Specifically, we first apply a text-based keyword extraction system to find a set of keywords from video transcripts. Then we apply various audiovisual content analysis techniques to identify cue contexts in which domain-specific keywords are more likely to appear. Finally, we adjust the keyword salience by fusing the audio, visual and text cues together, and "discover" a set of salient keywords.

Professionally produced educational or instructional

[1] For example, see http://video.google.com and http://video.yahoo.com

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 109–112,
New York, June 2006.

videos are the main focus of this work since they are playing increasingly important roles in people's daily lives. For the system evaluation, we used training and education videos that are freely downloadable from various DHS (Department of Homeland Security) web sites. These were selected because 1) DHS has an increasing need for quickly browsing, searching and re-purposing its learning resources across its over twenty diverse agencies; 2) most DHS videos contain closed captions in compliance with federal accessibility requirements such as Section 508.

2 A Text-based Keyword Extraction System

This section describes the text-based keyword extraction system, *GlossEx*, which we developed in our earlier work (Park et al, 2002). *GlossEx* applies a hybrid method, which exploits both linguistic and statistical knowledge, to extract domain-specific keywords in a document collection. *GlossEx* has been successfully used in large-scale text analysis applications such as document authoring and indexing, back-of-book indexing, and contact center data analysis.

An overall outline of the algorithm is given below. First, the algorithm identifies candidate glossary items by using syntactic grammars as well as a set of entity recognizers. To extract more cohesive and domain-specific glossary items, it then conducts pre-nominal modifier filtering and various glossary item normalization techniques such as associating abbreviations with their full forms, and misspellings or alternative spellings with their canonical spellings. Finally, the glossary items are ranked based on their confidence values.

The confidence value of a term $T, C(T)$, is defined as

$$C(T) = \alpha * TD(T) + \beta * TC(T) \quad (1)$$

where TD and TC denote the term domain-specificity and term cohesion, respectively. α and β are two weights which sum up to 1. The domain specificity is further defined as

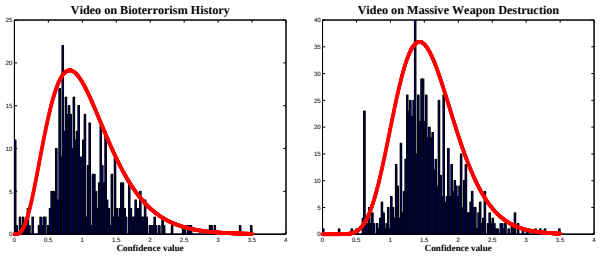
$$TD = \frac{\sum_{w_i \in T} \frac{P_d(w_i)}{P_g(w_i)}}{\mid T \mid} \quad (2)$$

where, $\mid T \mid$ is the number of words in term T, $p_d(w_i)$ is the probability of word w_i in a domain document collection, and $p_g(w_i)$ is the probability of word w_i in a general document collection. And the term cohesion is defined as

$$TC = \frac{\mid T \mid \times f(T) \times log_{10} f(T)}{\sum_{w_i \in T} f(w_i)} \quad (3)$$

where, $f(T)$ is the frequency of term T, and $f(w_i)$ is the frequency of a component word w_i.

Finally, *GlossEx* normalizes the term confidence values to the range of $[0, 3.5]$. Figure 1 shows the normalized distributions of keyword confidence values that we obtained from two instructional videos by analyzing their text transcripts with *GlossEx*. Superimposed on each plot is the probability density function (PDF) of a gamma distribution ($Gamma(\alpha, \gamma)$) whose two parameters are directly computed from the confidence values. As we can see, the gamma PDF fits very well with the data distribution. This observation has also been confirmed by other test videos.



(a) (b)

Figure 1: Normalized distribution of keyword saliencies for two DHS video, superimposed by Gamma PDFs.

3 Salient Keyword Extraction for Instructional Videos

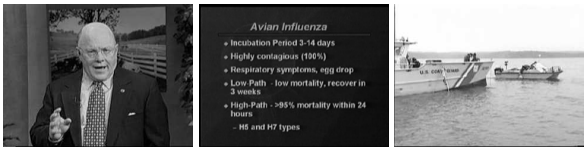
In this section, we elaborate on our approach for extracting salient keywords from instructional videos based on the exploitation of audiovisual and text cues.

3.1 Characteristics of Instructional Videos

Compared to general videos, professionally produced instructional videos are usually better structured, that is, they generally contain well organized topics and sub-topics due to education nature. In fact, there are certain types of production patterns that could be observed from these videos. For instance, at the very beginning section of the video, a host will usually give an overview of the main topics (as well as a list of sub-topics) that are to be discussed throughout the video. Then each individual topic or sub-topic is sequentially presented following a pre-designed order. When one topic is completed, some informational credit pages will be (optionally) displayed, followed by either some informational title pages showing the next topic, or a host introduction. A relatively long interval of music or silence that accompanies this transitional period could usually be observed in this case.

To effectively deliver the topics or materials to an audience, the video producers usually apply the following types of content presentation forms: host narration, interviews and site reports, presentation slides and information bulletins, as well as assisted content that are related with the topic under discussion. For convenience, we call the last two types as *informative text* and *linkage scene*

in this work. Figure 2 shows the individual examples of video frames that contain narrator, informative text, and the linkage scene.



(a) (b) (c)

Figure 2: Three visual content types: (a) narrator, (b) informative text, and (c) linkage scene.

3.2 AudioVisual Content Analysis

This section describes our approach on mining the aforementioned content structure and patterns for instructional videos based on the analysis of both audio and visual information. Specifically, given an instructional video, we first apply an audio classification module to partition its audio track into homogeneous audio segments. Each segment is then tagged with one of the following five sound labels: speech, silence, music, environmental sound, and speech with music (Li and Dorai, 2004). The support vector machine technique is applied for this purpose.

Meanwhile, a homogeneous video segmentation process is performed which partitions the video into a series of video segments in which each segment contains content in the same physical setting. Two groups of visual features are then extracted from each segment so as to further derive its content type. Specifically, features regarding the presence of human faces are first extracted using a face detector, and these are subsequently applied to determine if the segment contains a narrator. The other feature group contains features regarding detected text blobs and sentences from the video's text overlays. This information is mainly applied to determine if the segment contains informative text. Finally, we label segments that do not contain narrators or informative text as linkage scenes. These could be an outdoor landscape, a field demonstration or indoor classroom overview. More details on this part are presented in (Li and Dorai, 2005).

The audio and visual analysis results are then integrated together to essentially assign a semantic audiovisual label to each video segment. Specifically, given a segment, we first identify its major audio type by finding the one that lasts the longest. Then, the audio and visual labels are integrated in a straightforward way to reveal its semantics. For instance, if the segment contains a narrator while its major audio type is music, it will be tagged as *narrator with music playing*. A total of fifteen possible constructs is thus generated, coming from the combination of three visual labels (narrator, informative text and linkage scene) and five sound labels (speech, silence, music, environmental sound, and speech with music).

3.3 AudioVisual and Text Cues for Salient Keyword Extraction

Having acquired video content structure and segment content types, we now extract important audiovisual cues that imply the existence of salient keywords. Specifically, we observe that topic-specific keywords are more likely appearing in the following scenarios (a.k.a *cue context*): 1) the first N_1 sentences of segments that contain narrator presentation (*i.e.* narrator with speech), or informative text with voice-over; 2) the first N_2 sentences of a new speaker (*i.e.* after a speaker change); 3) the question sentence; 4) the first N_2 sentences right after the question (i.e. the corresponding answer); and 5) the first N_2 sentences following the segments that contain silence, or informative text with music. Specifically, the first 4 cues conform with our intuition that important content subjects are more likely to be mentioned at the beginning part of narration, presentation, answers, as well as in questions; while the last cue corresponds to the transitional period between topics. Here, N_1 is a threshold which will be automatically adjusted for each segment during the process. Specifically, we set N_1 to $min(\text{SS}, 3)$ where SS is the number of sentences that are overlapped with each segment. In contrast, N_2 is fixed to 2 for this work as it is only associated with sentences.

Note that currently we identify the speaker changes and question sentences by locating the signature characters (such as ">>" and "?") in the transcript. However, when this information is unavailable, numerous existing techniques on speaker change detection and prosody analysis could be applied to accomplish the task (Chen et al., 1998).

3.4 Keyword Salience Adjustment

Now, given each keyword (K) obtained from *GlossEx*, we recalculate its salience by considering the following three factors: 1) its original confidence value assigned by *GlossEx* ($C_{GlossEx}(K)$); 2) the frequency of the keyword occurring in the aforementioned cue context ($F_{cue}(K)$); and 3) the number of component words in the keyword ($|K|$). Specifically, we give more weight or incentive ($I(K)$) to keywords that are originally of high confidence, appear more frequently in cue contexts, and have multiple component words. Note that if keyword K does not appear in any cue contexts, its incentive value will be zero.

Figure 3 shows the detailed incentive calculation steps. Here, $mode$ and σ denote the mode and standard deviation derived from the *GlossEx* 's confidence value distribution. $MAX\_CONFIDENCE$ is the maximum confidence value used for normalization by *GlossEx*, which is set to 3.5 in this work. As we can see, the three aforementioned factors have been re-transformed into $C(K)$, $F(K)$ and $L(K)$, respectively. Please also note that we

have re-adjusted the frequency of keyword K in the cue context if it is larger than 10. This intends to reduce the biased influence of a high frequency. Finally, we add a small value ϵ to $|K|$ and F_{cue} respectively in order to avoid zero values for $F(K)$ and $L(K)$. Now, we have similar value scales for $F(K)$ and $L(K)$ ($[1.09, 2.xx]$) and $C(K)$ ($[0, 2.yy]$), which is desirable.

As the last step, we boost keyword K's original salience $C_{GlossEx}(K)$ by $I(K)$.

if $(C_{GlossEx}(K) >= mode$
$C(K) = \frac{C_{GlossEx}(K)}{mode}$
else $C(K) = \frac{C_{GlossEx}(K)}{MAX\_CONFIDENCE}$

if ($F_{cue}(K) > 10$)
$F_{cue}(K) = 10 + \log_{10}(F_{cue}(K) - 10)$
$F(K) = ln(F_{cue}(K) + \epsilon)$

$L(K) = ln(|K| + \epsilon)$

$I(K) = \sigma \times C(K) \times F(K) \times L(K)$

Figure 3: Steps for computing incentive value for keyword K appearing in cue context

4 Experimental Results

Four DHS videos were used in the experiment, which contain diverse topics ranging from bio-terrorism history, weapons of mass destruction, to school preparation for terrorism. The video length also varies a lot from 30 minutes to 2 hours. Each video also contains a variety of sub-topics. Video transcripts were acquired by extracting the closed captions with our own application.

To evaluate system performance, we compare the keywords generated from our system against the human-generated gold standard. Note that for this experiment, we only consider nouns and noun phrases as keywords. To collect the ground truth, we invited a few human evaluators, showed them the four test videos, and presented them with all candidate keywords extracted by *GlossEx*. We then asked them to label all keywords that they considered to be domain-specific, which is guidelined by the following question: "*would you be satisfied if you get this video when you use this keyword as a search term*?".

Table 1 shows the number of candidate keywords and manually labeled salient keywords for all four test videos. As we can see, approximately 50% of candidate keywords were judged to be domain-specific by humans. Based on this observation, we selected the top 50% of highly ranked keywords based on the adjusted salience, and examined their presence in the pool of salient keywords for each video. As a result, an average of 82% of salient keywords were identified within these top 50% of re-ranked keywords. In addition, the audiovisual cues improve precision and recall by 1.1% and 1.5% respectively.

videos	v_1	v_2	v_3	v_4
no. of candidate keywords	477	934	1303	870
no. of salient keywords	253	370	665	363
ratio of salient keywords	53%	40%	51%	42%

Table 1: The number of candidate and manually labeled salient keywords in the four test videos

5 Conclusion and Future Work

We described a mutimodal feature-based system for extracting salient keywords from instructional videos. The system utilizes a richer set of information cues which not only include linguistic and statistical knowledge but also sound classes and characteristic visual content types that are available to videos. Experiments conducted on the DHS videos have shown that incorporating multimodal features for extracting salient keywords from videos is useful.

Currently, we are performing more sophisticated experiments on different ways to exploit additional audio-visual cues. There is also room for improving the calculation of the incentive values of keywords. Our next plan is to conduct an extensive comparison between *GlossEx* and the proposed scheme.

References

Y. Park, R. Byrd and B. Boguraev. 2002. *Automatic Glossary Extraction: Beyond Terminology Identification*. Proc. of the 19th International Conf. on Computational Linguistics (COLING02), pp 772–778.

Y. Li and C. Dorai. 2004 *SVM-based Audio Classification for Instructional Video Analysis*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04).

Y. Li and C. Dorai. 2005 *Video frame identification for learning media content understanding*. IEEE International Conference on Multimedia & Expo (ICME'05).

M. Smith and T. Kanade. 1997 *Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques*. IEEE Computer Vision and Pattern Recognition, pp. 775-781.

G. Salton and J. McGill 1983. *Introduction to modern information Retrieval*. . New York: McGraw-Hill.

G. Salton and C. Buckley 1988. *Term-Weighting Approaches in Automatic Text Retrieval*. Information Processing & Management, 24 (5), 513-523.

S. Chen and P. Gopalakrishnan 1998. *Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion*. Proc. of DARPA Broadcast News Transcription and Understanding Workshop.