

# Leveraging Multi-modal Analyses and Online Knowledge Base for Video Aboutness Generation

Raj Kumar Gupta<sup>(✉)</sup> and Yang Yinping

Institute of High Performance Computing, Agency for Science,  
Technology and Research (A\*STAR), Singapore, Singapore  
{gupta-rk,yangyp}@ihpc.a-star.edu.sg

**Abstract.** The Internet has a huge volume of unlabeled videos from diverse sources, making it difficult for video providers to organize and for viewers to consume the content. This paper defines the problem of video aboutness generation (i.e., the automatic generation of a concise natural-language description about a video) and characterizes its differences from closely related problems such as video summarization and video caption. We then made an attempt to provide a solution to this problem. Our proposed system exploits multi-modal analyses of audio, text and visual content of the video and leverages the Internet to identify a top-matched aboutness description. Through an exploratory study involving human judges evaluating a variety of test videos, we found support of the proposed approach.

## 1 Introduction

Video sharing sites like YouTube and Vimeo have changed the way people consume, share and even produce multimedia content. The content on these sites is extremely diverse in their nature, ranging from news, weather forecast, sales demonstrations, talk shows, music, drama, as well as user self-recorded clips. Owing to the recent advancement of network infrastructure and growing popularity of social media, the growth of these unlabeled and misplaced videos has been phenomenal in its speed and volume, imposing a realistic challenge for video sharing sites to organize and for viewers to consume the content effectively. Thus, automatic extraction of useful descriptions of such videos is potentially of high value for practical deployment.

The generation of natural language descriptions of videos is receiving growing research interest recently [1–5]. Li et al. [3], for example, examined a deep convolutional neural networks based method to extract visual features from randomly selected video frames that were fed into recurrent neural networks to generate sentence description for each of these frames. The most relevant video description was then obtained by ranking the frame sentence descriptions using sentence-sequence graph. Thomason et al. [5] used visual recognition systems based on histogram of gradients, histograms of optical flow and motion boundary histogram features to predict high-level visual details such as the objects,

activities and scene present in a video. They then applied a factor graph model to integrate this visual information to select the best subject-verb-object-place description of a video.

Despite the increasing attention, existing methods reported to date focus on the generation of captions of short-duration, single-activity video clips [2, 3, 5]. To the best of our knowledge, little has been explored on methods that are capable of generating natural language descriptions of *more complex, longer-duration* videos such as those report situation development of an infectious disease outbreak or those illustrate an innovative hair styling procedure. We provide more discussion on this problem in Sect. 2.

We also noted that the methods developed to date revolve around extracting *visual* content from a video, yet the *audio* and *textual* content that may give additional information were not sufficiently examined. We explored the usefulness of both audio and text analysis techniques [6–8] on top of visual content extraction and generated more descriptors of the video. The multi-modal exploitation approach was shown to receive higher judgment ratings in the user evaluation study.

In terms of generating description using video global descriptors, existing works also tended to dive in grammatical sequencing of words to compose a description [3–5]. This “composition” approach can be extremely hard for complex videos as there can be a huge number of possibilities to compose a description for humans to appreciate. Here, we explored a “search” approach that uses the Internet (i.e., millions of titles from blogs and news that match the video’s global descriptors) as a knowledge base to algorithmically identify a top-matched aboutness description.

## 2 Video Aboutness: A Video Content Description in Concise Natural Language

In this paper, we consider the notion of “video aboutness” as a *concise* description *about* a video which needs to be informative, short and meaningful to a human viewer. The promise of a good video aboutness generation system is that aboutness description should be understood by new viewers and video sharing providers to quickly capture a central perspective of the video clip without watching it.

Automatic generation of video aboutness is a challenging new computing problem characterized in three dimensions. We characterized this problem in the following three dimensions.

First, at its basics, the video content description needs to be informative to answer the fundamental question: “What is about this video?”. In Library and Information Science, the term aboutness is introduced in the 1970s by Fairthorne [9] to express certain attributes of the text or document, i.e. what is said in a document. Similarly, in the Philosophy of Logic and Language, aboutness is understood as the way a piece of text relates to a subject matter or topic [10]. As such, the “aboutness” is fundamentally a description of the video.

Second, in the practical context of today’s social-technological landscape, the video description also needs to be short to read to a human user in the era of information overloading and fast content consumption.

Third, the description also needs to be considered good enough when assessed by viewers. This is in connection with the second form of aboutness that Fairthorne [9] distinguished: “intentional aboutness” referring to how the author views and intent to make a document is about, and “extensional aboutness” referring to how the document is reflected semantically.

The problem of video aboutness generation is different from two related tasks in visual computing and multimedia research. It differs from “video summarization” as the latter is essentially a task to identify the key frames or key events from the video that enables the viewers to gain maximum information about the target video in the minimum time [11]. Video aboutness is more closely related to “video caption”, but video caption methods are typically focused on composing descriptions of short video clips of 10–25 s in duration and consisting of a single activity (e.g., playing a piano, a person is running a race) [2, 3, 5]. In our work, we target on more naturally-occurring and complex videos (e.g., videos that report situation development of an infectious disease outbreak or that illustrate an innovative hair styling procedure). Such videos typically last more than two minutes and consist of multiple activities.

### 3 A Video Aboutness Generation System

We next describe a fully automatic system which is capable of generating the aboutness of a video. The proposed system consists of two major procedures, (1) the generation of global descriptors of the video using multiple sources of content information (audio, text and image processing), and (2) the generation of the final aboutness description leveraging output of (1) and the Internet as the knowledge base.

#### 3.1 System Workflow

Figure 1 provides the workflow of the system from application point of view: (a) A query video is entered as an input to the system, (b) The video is processed in four simultaneous procedures, including Audio Classification, Audio Keywords Detection, Textual Keywords Detection, and Image Information Extraction, to produce a rich set of global descriptors as intermediate outputs, (c) The global descriptors are used as keywords to retrieve news and blog articles, and are subsequently used to re-rank the retrieved news and blog articles based on similarity, (d) The title of top-ranked news or blog article is returned as the final output of the system.

#### 3.2 Global Description Generation

**Audio Classification.** First, we consider a classic audio analysis technique that generates different categorical information. This involves training of different audio classifiers such as speech, music, comedy (using the laughing sequences

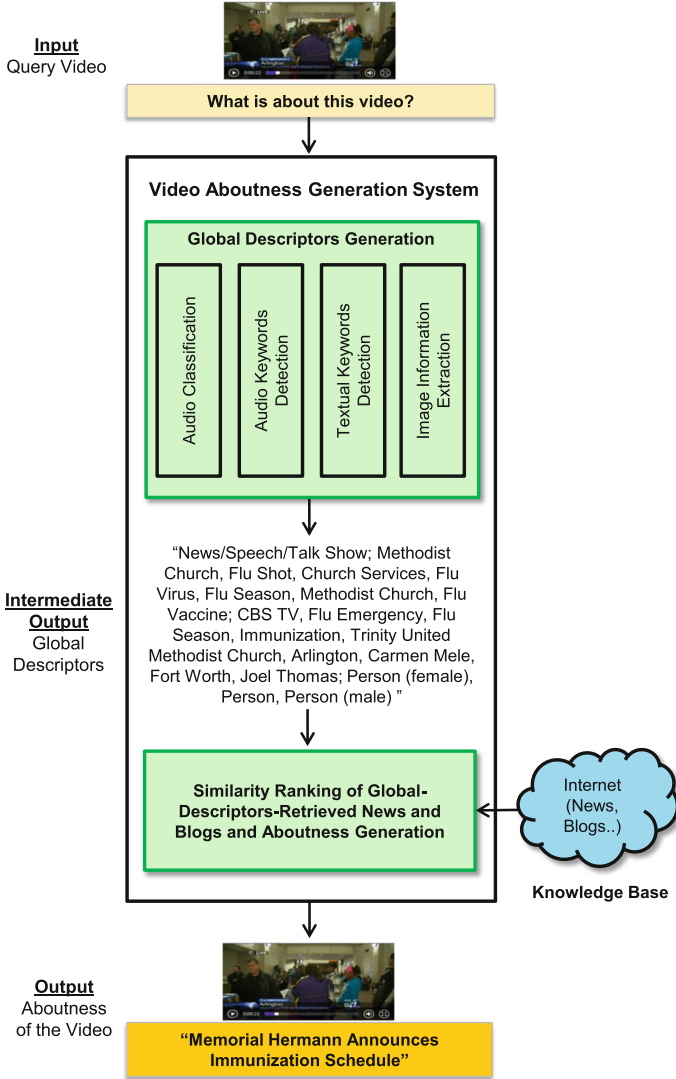


Fig. 1. The system workflow

from sitcoms/stand-up comedy videos), sports (using stadium noise and cheering sequences) to detect the genre of the video. Here, we extract the Mel-frequency cepstral coefficients features [7] (block size = 1024 and step size = 512) from the audio stream and use linear support vector machine [12] to train these classifiers. Therefore, these categories information forms the first part of the global descriptors.

In the example query video illustrated in Fig. 1, this procedure returns *News/Speech/Talk Show*.

**Audio Keywords Detection.** Second, because a high-level category description may not offer sufficient level of specifics and details about a video, we consider the audio stream of videos which gives the topic level information about a video. Here, we use CMUSphinx<sup>1</sup> to transcribe the audio stream into a textual transcript. Due to the noise in audio streams, these direct outputs of the textual transcripts are not very accurate.

To enhance the speech-to-text performance, we fetch the keywords in the form of tags with news and blogs articles available on Internet. These keywords, often containing unigrams, bigrams and trigrams, are then used to rebuild the language model<sup>2</sup>, such that this language model is subsequently used to transcribe the audio stream. After transcribing the audio stream into a textual transcript, all available keywords from the transcript can be extracted as detected keywords to form the second part of the intermediate global descriptors.

In the example query video illustrated in Fig. 1, this procedure returns *Methodist Church, Flu Shot, Church Services, Flu Virus, Flu Season, Methodist Church, Flu Vaccine*.

**Textual Keywords Detection.** Besides audio content, online videos often contain textual information such as sub-titles of news. To obtain the textual information from the video, we extract the video frames in every 5s. Here, we use OpenCV implementation of the method proposed by Neumann and Matas [8] for text localization within each extracted frame and use tesseract-ocr<sup>3</sup> to extract the text from each of these localized image regions. In this image derived text, we further search for keywords that have been used to train the language model described in Audio Keywords Detection.

In addition, we also extract the valid names from the image text. To capture the names from the image text, we used the first name and last name dataset available on the United States Census Bureau website<sup>4</sup>. These textual content derived keywords then form the third part of the intermediate global descriptors.

In the example query video illustrated in Fig. 1, this procedure returns *CBS TV, Flu Emergency, Flu Season, Immunization, Trinity United Methodist Church, Arlington, Carmen Mele, Fort Worth, Joel Thomas*.

**Visual Content Extraction.** Lastly, to obtain the image content from visual stream, we extract all the shots using FFmpeg<sup>5</sup>. From each of these shots, we select a representative image based on its visual attributes (e.g. sharpness, lighting) to perform image analysis. We then use a few latest computer vision techniques to evaluate each of these images to identify the objects (e.g. cars, horse, motorbike) [13], people with attributes (e.g. male or female, ethnicity) [14, 15],

<sup>1</sup> <http://cmusphinx.sourceforge.net/>.

<sup>2</sup> <http://www.speech.cs.cmu.edu/tools/lmtool-new.html>.

<sup>3</sup> <https://github.com/tesseract-ocr>.

<sup>4</sup> [http://www.census.gov/topics/population/genealogy/data/1990\\_census/1990\\_census\\_namefiles.html](http://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html).

<sup>5</sup> <https://www.ffmpeg.org/>.

stuff (e.g. water, grass, sky) [14, 18], and indoor (e.g. bathroom, kitchen) [16, 17] and outdoor scenes (e.g. beach, highways) [17–19]. This generates specialized and more fine-grained keywords that form the fourth part of the intermediate global descriptors.

In the example query video illustrated in Fig. 1, this procedure returns *Person (female)*, *Person*, *Person (male)*.

### 3.3 Similarity Ranking Using Internet as Knowledge Base

After this rich set of global descriptors is obtained, these global descriptors are used as keywords to retrieve the news and blog articles from the Internet. These retrieved news and blog articles are re-ranked based on their similarity scores with the global description of the video. The similarity scores are computed based on the frequencies of the global video descriptions in the retrieved articles.

Finally, the title of the top ranked article (e.g., *Memorial Hermann Announces Immunization Schedule* as in the example query video in illustrated Fig. 1) is used to describe the input video.

## 4 Experiment

### 4.1 Data and System Processing

Existing methods are typically evaluated on video contents constrained within a small set of known objects and single action activities (e.g. two teams playing football) [2, 3, 5]. We are interested in examining a wider variety of videos, which are unconstrained with objects and activities content.


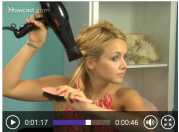

To assess our system, we downloaded a total of 21 test videos<sup>6</sup> covering a wide variety of content including news (videos 1–3, 8, 9, 11, 15, 21), skills illustrations (videos 4–6), sales demonstrations (videos 7, 12), talk shows (10, 13), weather forecast (video 14), self-recorded clip (video 16), and music (videos 17–20). The duration of the videos ranges from 0.24 to 5.20 min, averaged at 2.09 min. These videos were processed in a system implemented based on Sect. 3.

Table 1 present three examples of the video aboutness results<sup>7</sup> generated using the proposed approach. It also shows the global descriptions of the videos that have been extracted after audio and visual analysis that are intermediate outputs of the system. Apparently, it can be seen that the global descriptions of these videos are very coarse on their own, and the final aboutness outputs are shorter and more informative.

<sup>6</sup> The dataset can be downloaded from [https://www.dropbox.com/sh/315lz0r7i552kjg/AADCu1wr\\_NLdVau79kvPVEXLa](https://www.dropbox.com/sh/315lz0r7i552kjg/AADCu1wr_NLdVau79kvPVEXLa).

<sup>7</sup> The results of all test videos can be downloaded from [https://www.dropbox.com/s/c1e126dps0brd0r/aboutness\\_results.pdf](https://www.dropbox.com/s/c1e126dps0brd0r/aboutness_results.pdf).

**Table 1.** Video aboutness - input videos, intermediate output and final output (three examples)

ID	Input Video	Intermediate Output - Global Descriptors		Video Aboutness	
		Visual Only	Multi-modal	Visual Only	Multi-modal
3.		Arkansas, General Assembly, Immigration, Lottery Scholarships, School Funding, Craig Cannon; Person (male), person (female), building	News/Speech/Talk Show; General Assembly, Lawmakers, State Capital, Public Health, Federal Government, Republican, Medicate Program, General Assembly, State Capital, Lawmakers, Illegal Immigrants; Arkansas, General Assembly, Immigration, Lottery Scholarships, School Funding, Craig Cannon; Person (male), person (female), building	High School - Arkansas Department Of Higher Education	State Legislative Tracker: Lawmakers Dive Into 2013 Business
5.		Hair Clip, Hand-Held Dryer, Howcast, Bristle; Person (female)	Speech+Music/Talk Show; Heat Protective; Hair Clip, Hand-Held Dryer, Howcast, Bristle; Person (female)	How To Blow Dry Your Hair Straight	How To Blow Dry Your Hair Straight
19.		The Goo Goo Dolls; Person, TV	Speech+Music/Talk Show; Karaoke Show; The Goo Goo Dolls; Person, TV	Goo Goo Dolls	Iris By The Goo Goo Dolls (Karaoke)

## 4.2 User Study

To assess the quality of the system efficacy from a viewer’s perspective, we invited eleven human judges to evaluate the generated aboutness descriptions. These judges, four females, are researchers employed by a large publicly funded research agency, having various backgrounds in computer science, information systems, physics, and psychology. Their minimum education level is Bachelor degree, and a majority (nine) holds Ph.D. degrees.

The judges were instructed to watch each of the videos, understand the content and then rate the content descriptions with a score that best describes their evaluation on each description. The scores are in a range of 1 to 7, where

- 1 denotes “*This is a very poor description about the video*”
- 4 denotes “*This is a fair description about the video*”
- 7 denotes “*This is a very good description about the video*”

Two video aboutness descriptions were simultaneously presented to the judges, one derived from using *visual only* global descriptors and another derived from *multi-modal* (i.e., audio, text and visual) global descriptors. The two descriptions were randomly positioned to reduce order effect.

The judges were given no time limit to do the evaluation and performed the evaluations of all the 21 videos independently. Figure 2 shows a sample interface of the user study site<sup>8</sup>.



Fig. 2. Sample user study interface

Results show that the visual-only based video aboutness descriptions received an average score of 4.1 (*std. dev.* = 0.7; *median score* = 4.1) from the judges. The multi-modal-based video aboutness descriptions received an average score of 4.6 (*std. dev.* = 0.6; *median score* = 4.9) from the judges, supporting the additional usefulness of the audio processing components in our proposed system. A paired sample *t*-test showed that the multi-modal vs. visual-only processing difference is statistically different ( $t = 2.23, p = 0.01$ ). This indicates that the proposed automatic video aboutness approach has a fair level of efficacy as evaluated by the independent human judges on a variety of videos.

As a post-hoc analysis, we further examined the types of videos that received lower-than-average scores. Results show that aboutness descriptions of self-recorded clip (video 16) and music videos with commentaries (videos 17, 19) generally received lowest scores, in that if one only considers the remaining videos, the visual-only based descriptions received an average score of 4.5 (*std. dev.* = 0.7; *median score* = 4.7) and the multi-modal-based descriptions received an average score of 5.0 (*std. dev.* = 0.6; *median score* = 5.2)

<sup>8</sup> The full user study site can be accessed from here: <http://52.76.48.244/userstudy/>.



( $t = 2.23$ ,  $p = 0.00$ ). This result suggested useful pointers for future extensions, such as incorporating activity detection [20] and eliciting emotional descriptors (e.g., frustrated, sad, lonely, joyful) [21, 22], that may help to provide even richer global descriptors.

## 5 Conclusion

This paper is motivated to provide a feasible solution to a new challenge of automatically generating informative, short and meaningful descriptions of unlabeled online videos. We characterized the problem of video aboutness generation and described a system leveraging various latest multi-modal data processing techniques in conjunction with an innovative use of the Internet as a knowledge base. Experimental results supported the benefit of multi-modal analyses and offered support that the system can generate reasonably well aboutness descriptions for a wide variety of online target videos. In future work, it would be interesting to further exploit latest visual computing techniques such as activity detection and emotion detection to enrich the global descriptors and thereby enhance the quality of aboutness generation.

**Acknowledgement.** This research is supported by the Social Technologies + Programme funded by A\*STAR Joint Council Office. We thank Tong Joo Chuan for the encouragement to pursue this research direction and are grateful to the volunteers who participated in the user study.

## References

1. Barbu, A., Bridge, E., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangquan, J., Siskind, J.M., Waggoner, J., Wang, S., Wei, J., Yin, Y., Zhang, Z.: Video in sentences out. In: Association for Uncertainty in Artificial Intelligence (UAI) (2012)
2. Khan, M.U.G., Gotoh, Y.: Describing video contents in natural language. In: Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (2012)
3. Li, G., Ma, S., Han, Y.: Summarization-based video caption via deep neural networks. In: ACM Multimedia (2015)
4. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. In: International Conference on Computer Vision (2013)
5. Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.: Integrating language and vision to generate natural language descriptions of videos in the wild. In: 25th International Conference on Computational Linguistics (COLING) (2014)
6. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnick, A.I.: Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: International Conference on Acoustics Speech and Signal Processing (2006)

7. McKinney, M.F., Breebaart, J.: Features for audio and music classification. In: International Conference on Music Information Retrieval (2003)
8. Neumann, L., Matas, J.: Text localization in real-world images using efficiently pruned exhaustive search. In: International Conference on Document Analysis and Recognition (2011)
9. Fairthorne, R.A.: Content analysis, specification and control. *Ann. Rev. Inf. Sci. Technol.* **4**, 73–109 (1969)
10. Searle, J.: *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, Cambridge (1983)
11. Khosla, A., Hamid, R., Lin, C.J., Sundareshan, N.: Large-scale video summarization using web-image priors. In: Computer Vision and Pattern Recognition (2009)
12. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **27**(1–27), 27 (2011)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
14. Aghajanian, J., Warrell, J., Prince, S.J., Li, P., Rohn, J.L., Baum, B.: Patch-based within-object classification. In: International Conference on Computer Vision (2009)
15. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645 (2010)
16. Hays, J., Efros, A.A.: Scene completion using millions of photographs. In: ACM SIGGRAPH (2007)
17. Trefny, J., Matas, J.: Extended set of local binary patterns for rapid object detection. In: Proceedings of the Computer Vision Winter Workshop (2010)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
19. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
20. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
21. Schwarz, N., Clore, G.L.: Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513–523 (1983)
22. Schwarz, N.: Feelings-as-information theory. In: Van Lange, P., Kruglanski, A., Higgins, E.T. (eds.) *Handbook of Theories of Social Psychology*, pp. 289–308 (2012)