論文 / 著書情報
Article / Book Information

Title	Speech Technology Plays a Key Role in Video Semantic Indexing
Author	Koichi Shinoda
Citation(English)	First International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA) at ACM Multimedia 2012, , , pp. 1-2
Issue date	2012, 10
Copyright	Copyright (c) 2012 Association for Computing Machinery
Set statement	Copyright (C)2012 Association for Computing Machinery(ACM), . This is the author's version of the work. It is posted here by personal use. Not for redistribution. The definitive version of Record was published in Koichi Shinoda, First International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA) at ACM Multimedia 2012,2012,pp. 1-2.10.1145/2390214.2390216
Note	このファイルは著者（最終）版です。 This file is author (final) version

Speech Technology Plays a Key Role in Video Semantic Indexing

[Extended Abstract]

Koichi Shinoda
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
shinoda@cs.titech.ac.jp

ABSTRACT

Video semantic indexing is a core task in content-based video retrieval (CBVR), in which a user submits a text query for an object or a scene to a search system and the system returns video shots that include the object or scene. We introduce an emerging framework for this task, which heavily relies on statistical speaker verification and adaptation techniques. It employs Gaussian-mixture-model (GMM) supervectors and support vector machines (SVM) to detect a large variety of objects and scenes robustly from video. It has shown excellent performance in the Semantic indexing task of the TRECVID 2011 workshop, where a large archive of consumer-produced Internet videos are used for evaluation.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]: Video

General Terms

Algorithms

Keywords

video semantic indexing, speaker adaptation, speaker verification, Gaussian mixture model, support vector machine

1. SEMANTIC INDEXING (SIN)

National Institute of Standards and Technology (NIST) in the US has held the TREC Video Retrieval Evaluation (TRECVID) workshop every year since 2001 to promote content-based video retrieval (CBVR) research and development [1]. Many research organizations participate in this workshop, and compete with each other for several CBVR tasks. Their methods and results are open to the public on the TRECVID web page [1], which is a showcase of the state-of-the-art CBVR technologies.

The most important task in TRECVID is Semantic INdexing (SIN). It has been conducted for the past ten years starting in 2002 and has had the largest number of participants amongst the various TRECVID tasks. In this task, a query comes in the form of a word or a phrase called a concept such as “desk”, “night scape”, and “dancing”. A search system should find shots including the concept from a large archive of Internet consumer video clips. Here a shot is a sequence of video frames between camera shot changes, whose duration is typically 10− 30 seconds. Figure 1) shows examples of detected shots. This task corresponds to Spoken Term Detection (STD) in the study of speech document search.

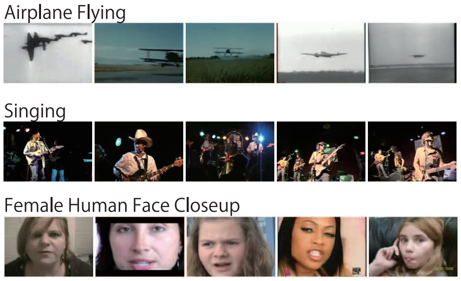

Figure 1: SIN top-5 results for three concepts [4].

Most SIN methods have been based on the Bag of Visual Words (BoW) framework, which is imported from generic object detection techniques from still images in the image recognition field. In this framework, features such as scale invariant feature transform (SIFT) [3] are extracted from video frames and they are clustered to form a codebook. For each shot, a code histogram is obtained by counting the number of occurrences of each code word. This code histogram, a vector with a dimension equal to the codebook size, is expected to represent the characteristics of the shot. Support vector machines (SVM) are often used to detect the shots which include the target concept.

Recently, however, many techniques have been proposed, which effectively utilize the characteristics of video data. Previously, only features extracted from a key frame of a shot were used. Nowadays, features from many frames are often used, and contribute to an increase in the robustness of the detectors against various dynamic changes within a shot. Audio features such as Mel-frequency cepstral coefficients (MFCC) significantly improved the detection performance for many concepts related to audio, such as “Infant” and “Car race”. Many systems now use the multi-kernel approach where several audio and visual features are extracted and one SVM is provided for each of them. The outputs from those SVMs are combined to obtain the detection score.

2. SPEECH TECHNOLOGY FOR SIN

The size of the TRECVID SIN task is rapidly increasing

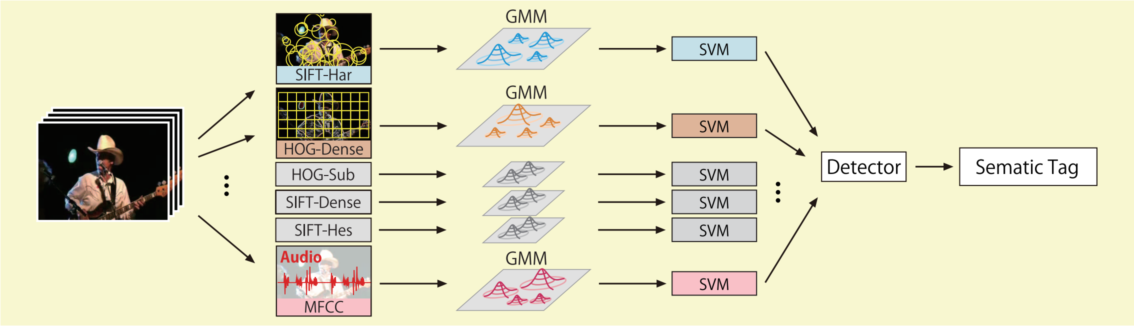

Figure 2: A video semantic indexing method [4] using six different features and providing one SVM for each.

year by year. In 2011, the total length of video data provided was 600 hours, which consisted of more than 400,000 shots. The number of concepts to be detected was 346. The quality of the data is rather low. Their variety is also large. Some of them are annotated with semantic labels but their quality is not always good. A larger model with more features is required, but the data insufficiency often deteriorates its performance. A system for a specific concept in a specific condition cannot be applied anymore. A generic system that is robust against various changes such as quality and data size has been strongly demanded.

This present situation of video semantic indexing reminds us of speech/speaker recognition in the 90's, 20 years ago. At that time, speech researchers faced exactly the same problems. The solution they found was a robust data-driven approach which heavily relied on probability theory. Thanks to the advancement of computation technology, the same approach is now ready to use for video semantic indexing which requires much more computational resources than the speech tasks of the 90's.

3. GMM SUPERVECTOR AND SVM

We introduce a video semantic indexing method [4] which heavily uses speech/speaker recognition technologies (Figure 2). It is an extension of the BoW framework to the probabilistic framework and had the highest detection performance in the TRECVID2011 SIN task.

In this method, a Gaussian mixture model (GMM) is provided for each shot. A GMM consists of more than one mixture component, each of which is a Gaussian distribution. Each mixture component corresponds to a code in BoW, its mean vector corresponds to a code vector of the code, and its weight can be regarded as the normalized occurrence count of the code in a code histogram. While only one code is assigned for one input in the BoW framework, one input belongs to many distributions with different weights in a shot GMM. This soft assignment mitigates the effect from the quantization error.

To achieve robustness against data sparseness, a model adaptation techniques from the speech field, maximum a posteriori (MAP) adaptation [5], is used to estimate the GMM mean vectors. In this technique, a universal background model (UBM) is first estimated from all training data and their parameters are effectively utilized as prior knowledge.

Next, a GMM-based speaker verification method [6] is used to obtain detection scores. A GMM supervector, which is made by concatenating all mean vectors in a GMM, is used as an input to a SVM. A simplified Fisher kernel is used as the SVM kernel.

Furthermore, to save the computational cost for the parameter estimation of each shot GMM, a fast search method using a tree-structured GMM is employed. This method successfully reduced the cost by 75% without any degradation in detection performance.

4. FEATURE DIRECTION

Video semantic indexing corresponds to isolated word recognition with a limited vocabulary size in speech research. Of course we, users, would like to search video clips that consist of multiple shots, by using multiple words. Video researchers have started to explore semantics in video concepts (e.g., [7]). TRECVID also defined a new task recently in 2010, which aims to extract a video clip including an event, such as "Getting a vehicle unstuck" from an Internet video archive. This direction again reminds speech researchers of their past history developing large vocabulary continuous speech recognition. Speech researchers can contribute a lot to video processing.

5. REFERENCES

[1] TREC Video Retrieval Evaluation. http://trecvid.nist.gov/

[2] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation Campaigns and TRECVid," In Proc. of ACM Multimedia MIR workshop, pp. 321-330, 2006.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[4] N. Inoue and K. Shinoda, "A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors," IEEE Trans. Multimedia, vol. 14, no. 4, pp. 1196-1205, 2012.

[5] J. L. Gauvain and C.-H Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, 1994.

[6] W. M. Campbell and D. A. Raynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, vol. 13, no. 5, pp. 308-311, 2006.

[7] M. Naphade, S.-F. Chang, A. Hauptmann, and J. Curtis, "Large scale concept ontology for multimedia," IEEE Multimedia, vol. 13, no. 3, pp. 86-91, 2006.