A Fast Video Event Recognition System and Its Application to Video Search

Yu-Gang Jiang[§], Qi Dai[§], Yingbin Zheng[§], Xiangyang Xue[§], Jie Liu[‡], Dong Wang[‡]
[§]School of Computer Science, Fudan University, Shanghai, China
[‡]Huawei Technologies, Beijing, China
{ygj,11210240046,ybzh,xyxue}@fudan.edu.cn,
{jane.liujie,dave.wangdong}@huawei.com

ABSTRACT

Techniques for recognizing complex events in diverse Internet videos are important in many applications. State-of-the-art video event recognition approaches normally involve modules that demand extensive computation, which prevents their application to large scale problems. In this demonstration, we present a fast video event recognition system, which requires just a few seconds to process a general YouTube video with a few minutes of duration. The development of this system is grounded on several important findings from a large set of empirical studies, where we systematically evaluated many technical options for each critical module of a present-day video event recognition framework. Pooling the insights gained from this study leads to a speeded-up event recognition system that is 220-times faster than a decent baseline while still has a high degree of recognition accuracy.

We also demonstrate the technical feasibility of using event recognition results as the sole clue for video search, where the similarity of videos is determined based on the consistency of the event recognition confidence scores. We showcase this capability using an Internet video dataset containing about 10 thousands of YouTube videos. Very promising results were observed.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation, Design.

Keywords

Fast video event recognition, Internet videos, speed efficiency, video search.

1. INTRODUCTION

With the explosive growth of videos on the Web, there is a strong need of automatic video content recognition solutions. Such techniques can be used in a wide range of applications such as personal video collection management, Web scale video search, smart advertising, etc. In this work, we are interested in recognizing high-level video events like "birthday party", "parade", "skiing" and so on, a problem that is receiving increasing research attention.

Although progress has been made in the past few years, the current video event recognition systems often involve modules that are extremely expensive to compute, such as the extraction of spatial-temporal interest points [3]. Different from the previous works which focused mostly on recognition accuracy, we strive to improve recognition speed while still maintain a good accuracy.

In this demonstration, we present a fast video event recognition system. On a regular laptop computer, only a few seconds are needed to process (including data preprocessing, feature extraction, classification and multimodal fusion) a video of a few minutes of duration. This is significantly faster than many existing systems. In addition to showing the speed efficiency of our event recognition system, we also demonstrate a potential application of event recognition in video search, where video similarity is estimated by the consistency of detected events. Although video search on the Web (e.g., YouTube search engine) is mostly executed by pure text matching, event-level content semantics offers a different and complementary view in measuring video similarity. In addition, for personal video collections or other special domain archives (e.g., documentary) where very few or no text annotations are available, event understanding provides a key clue for search.

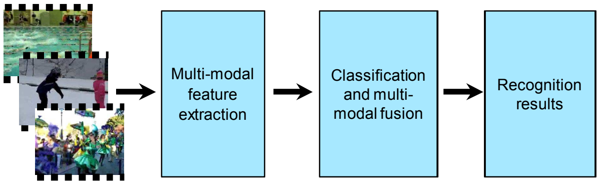

Multi-modal feature extraction
Classification and multi-modal fusion
Recognition results

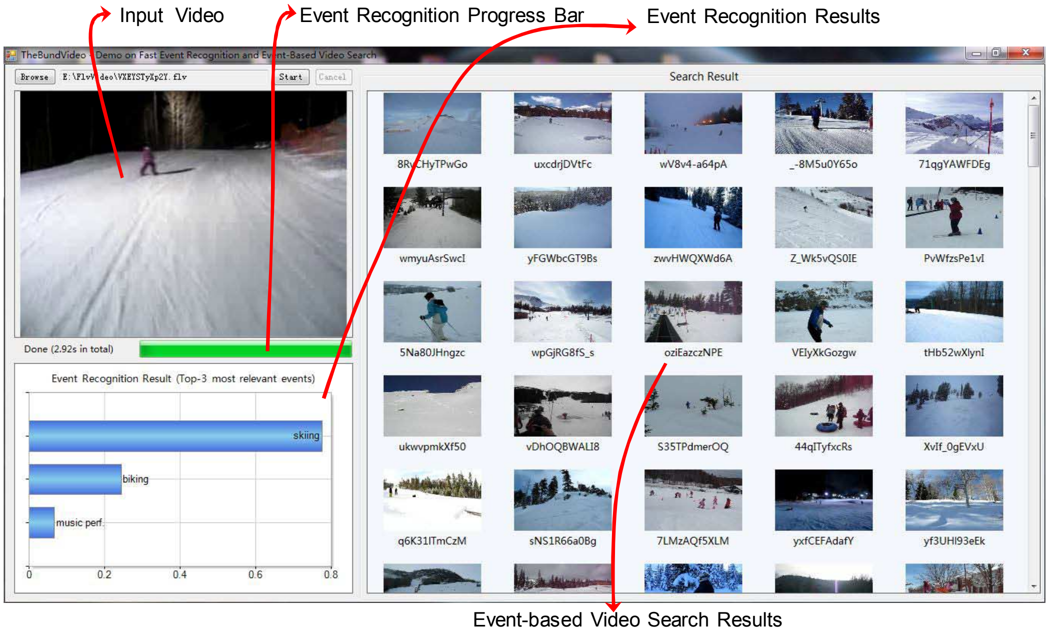
Figure 1: A typical video event recognition pipeline.

2. CORE TECHNIQUES

The proposed event recognition system follows a popular pipeline as shown in Figure 1, where the core techniques are identified based on a rigorous study conducted in [1]. We summarize several important findings and the techniques used in each module as follows.

1. Frame sampling: This is a preprocessing step, as using only a subset of the visual frames or audio clips (often called

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.



TheBundVideo - Demo on Fast Event Recognition and Event-Based Video Search

E:\FlvVideo\VXEYSTyXp2Y.flv

Done (2.92s in total)

Figure 2: The main interface of our demonstration system. Input (query) video is displayed in the upper-left corner, and event recognition results are shown in the lower-left corner (top-3 most relevant events; bar length indicates classifier confidence). Between the two there is a horizontal bar visualizing the recognition progress. Event-based video search results are shown on the right, where the videos are ordered according to their relatedness to the query, from left to right and top to bottom.

frames in the audio community) always saves feature extraction time. We evaluated the number of needed frames for both modalities, and found that sub-sampling audio frames always hurts the recognition accuracy but using a subset of visual frames (16 frames per video from our study) does not. Therefore we uniformly sampled 16 visual frames per video and kept the entire audio soundtrack for processing.

2. Feature: Through evaluating a large set of audio/visual features, we observed that dense SURF visual feature is good in both accuracy and speed. Combining it with MFCC audio feature leads to a substantial improvement of around 10% over SURF alone[1]. We also found that further adding expensive visual features like STIP does not contribute too much to recognition accuracy. Therefore dense SURF and MFCC features were adopted in this system. Both features were represented using the bag-of-words framework (cf. [1] for details).

3. Classification: We adopted SVM in the classification module and compared several kernel choices, including the widely used χ^2 kernel, histogram intersection (HI) kernel, and Maji's fast approximated HI kernel [4]. We observed that fast HI kernel is similar in accuracy to standard χ^2 and HI kernels, but is several hundreds of times faster, which was thus selected.

4. Multi-modal fusion: Early fusion and late fusion were evaluated. Early fusion, which concatenates representations from SURF and MFCC before classification, requires only one classifier and was also found to be slightly better than late fusion, which needs to run two classifiers separately. Early fusion was adopted.

Integrating all these findings leads to a system that can process an 80s video in just 4.56s on a regular laptop computer (including feature extraction and classification using 20 event models). This is similar in accuracy but is 220 times faster than a baseline system evaluated in [1], which produced very competitive results in the U.S. NIST TRECVID evaluation (Multimedia Event Detection Task[2]).

In our proof-of-concept system of event-based video search, we use the event recognition scores (i.e., 20-d vectors) as features and adopt Cosine similarity to measure the relatedness of two videos. The CCV dataset is used as the target database, which contains about 10 thousands of YouTube videos.

3. SYSTEM INTERFACE AND IMPLEMENTATION

Our system interface is shown and explained in Figure 2. Two capabilities are presented in this demo: 1) event recognition speed and accuracy, and 2) event-based video search. The system was implemented using mainly C++ on Windows 7 with Microsoft Visual Studio 2010. A few components (e.g., the fast HI kernel SVM [4]) were originally implemented in MATLAB, which were compiled into DLLs.

In summary, this demo showcases that high-level video events can be very efficiently recognized, and event recognition output can serve as a key clue for video search, which is especially useful for managing the ever-expanding personal video archive that does not have sufficient textual annotations. For future work, we will increase the number of event classes to the scale of several hundreds.

4. REFERENCES

[1] Y.-G. Jiang. SUPER: Towards real-time event recognition in Internet videos. In *Proc. of ACM ICMR*, 2012.

[2] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proc. of ACM ICMR*, 2011.

[3] I. Laptev. On space-time interest points. *IJCV*, 64:107–123, 2005.

[4] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. of IEEE CVPR*, 2008.

[1] Measured on Columbia Consumer Video (CCV) dataset [2].

[2] `http://www.nist.gov/itl/iad/mig/med.cfm`