Towards Textually Describing Complex Video Contents with Audio-Visual Concept Classifiers

Chun Chet Tan[†], Yu-Gang Jiang[‡], Chong-Wah Ngo[†]
[†]Dept of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[‡]School of Computer Science, Fudan University, Shanghai, China
cctan2@student.cityu.edu.hk, ygj@fudan.edu.cn, cwngo@cs.cityu.edu.hk

ABSTRACT

Automatically generating compact textual descriptions of complex video contents has wide applications. With the recent advancements in automatic audio-visual content recognition, in this paper we explore the technical feasibility of the challenging issue of precisely recounting video contents. Based on cutting-edge automatic recognition techniques, we start from classifying a variety of visual and audio concepts in video contents. According to the classification results, we apply simple rule-based methods to generate textual descriptions of video contents. Results are evaluated by conducting carefully designed user studies. We find that the state-of-the-art visual and audio concept classification, although far from perfect, is able to provide very useful clues indicating what is happening in the videos. Most users involved in the evaluation confirmed the informativeness of our machine-generated descriptions.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Textual descriptions of video content, audio-visual concept classification.

1. INTRODUCTION

Precise video content description is useful for many applications such as video search. Commercial search engines have been using some content-relevant texts, such as those extracted from captions/contexts of online videos and scripts of movies and TV series. However, these textual annotations provided by humans are far from adequate for enormous online videos which often neither have scripts nor meaningful captions. This motivates the need of developing automatic solutions for generating video content descriptions.

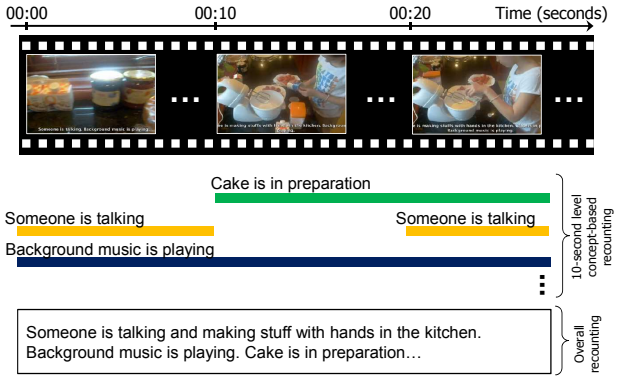


Figure 1: Video content recounting with audio-visual concept classifiers. Given a video (shown on the top), our proposed approach first classifies several audio-visual concepts, and then automatically produces textual descriptions based on the concept classification results.

In [3], an interesting method was proposed for textually describing videos. However, this approach was applied to simple data from constrained scenarios and is therefore infeasible to be implemented for real-world complex videos. Fortunately, in recent years there have been many research efforts devoted to visual and audio content recognition. For example, in [5], Laptev et al. proposed a very effective method based on spatial-temporal interest points for learning human actions in Hollywood movies. Jiang et al. [2] used a multimodal approach for complex video event recognition in Internet videos, and in [6], Lee et al. tried to identify semantic concepts in consumer videos by using audio feature alone. In each of the aforementioned works on content recognition, the authors have shown promising results on real-world video data. Although these content recognition techniques are still far below perfect, recent studies have found them helpful for many applications, such as video search [1, 9].

In this paper, we conduct a pilot study to explore the technical feasibility of automatically describing complex video contents. Our approach is grounded on the cutting-edge techniques for recognizing visual and audio semantics in videos. According to the recognition outputs, video contents are automatically described at both short clip level and

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

entire video level. The mapping from content recognition results to textual descriptions is done by simple rule-based methods. Using a set of Internet videos containing three complex event topics, we study to what extent the imperfect automatic audio-visual concept classifiers can contribute to textual video content recounting. Carefully designed user studies verify the promise of our idea. Figure 1 shows an example of content description output by our approach. In the following, we first present our proposed approach, and then describe experiments and discuss results.

2. VIDEO CONTENT RECOUNTING

Our proposed video content recounting framework consists of two major components: audio-visual concept learning and rule-based textual description generation.

2.1 Audio-Visual Concept Learning

Central to automatic content recounting is the recognition of semantic contents in videos. To this end, we develop automatic classifiers for identifying a set of audio/visual semantic concepts. We follow the state-of-the-art approaches in [5, 2, 6]. Specifically, we extract three kinds of audio-visual features: 2D static SIFT [7], 3D spatial-temporal interest points (STIPs) [4], and MFCC audio descriptors. For SIFT, we use two standard detectors DoG [7] and Hessian Affine [8], and each detected local image patch is described by a 128-dimentional gradient histogram. STIPs are detected at multiple spatial and temporal scales. Laptev's method [4] is adopted to locate keypoints with significant local variation in both space and time. Histogram of Oriented Gradients (HOG; 72 dimensions) and Histogram of Optical Flow (HOF; 90 dimensions) descriptors are computed for the detected STIPs. Eventually HOG and HOF descriptors are concatenated into a 162-dimensional vector for each STIP. In contrast to the two visual descriptors that are computed based on sparse detectors, the MFCC features are densely extracted in the audio track of the videos — we compute a 60-dimensional MFCC feature in every 32ms temporal window, and nearby windows have 16ms overlap.

Now that we have multiple SIFT, STIP, and MFCC descriptors extracted from each video. The popular bag-of-word representation is then applied to convert the 3 sets of descriptors separately into 3 fixed-dimensional feature vectors. We use hierarchical k-means to generate a visual vocabulary of 4,000 words for both STIP and MFCC, and two smaller vocabularies of 500 words for DoG-SIFT and Hessian-SIFT respectively. The choices of visual or audio vocabulary size are based on empirical evidences from the prior studies in [5, 2].

In audio-visual concept learning, we use each 10-second video clip, instead of the entire video, as a data sample, so that we can obtain finer-grained audio-visual clues for content recounting. For SIFT feature, two spatial layers (1×1 and 2×2) are used in the vector quantization process, producing a 5,000-dimensional feature vector for each 10-second clip ($1 \times 1 \times (2 \times 500) + 2 \times 2 \times (2 \times 500)$). For STIP and MFCC, no spatial/temporal partitioning is used.

With the three types of audio-visual features, we use SVM classifier for concept learning. Following [2], given an audio/visual concept, depending on the type of the concept, we empirically pick the best suitable feature to train a SVM classifier. The concepts used in our experiments and their training data will be introduced in the experiment section.

2.2 Concept-based Content Recounting

The second stage of our approach is to utilize the concept classification results to generate content recounting. For this, we take a simple rule-based approach by predefining a set of templates according to the concepts under consideration.

Before generating the textual descriptions, we first try to predict high-level events happening in the target videos based on audio-visual concept classification. The high-level event prediction can be used for alleviating the effect of noisy concept classification. We manually define a $k \times l$ event-concept relevancy matrix $\mathbf{R}$ for this purpose, where k is the number of events and l is the number of concepts. Since practically k is very likely to be in the scale of a few hundreds, it is feasible to let humans fill in this relevancy matrix. The entries of the matrix are either +1 for concepts with strong positive correlations to the event, or -1 for those with negative correlations to the event. For example, concept "outdoor" is assigned with a value -1 while "kitchen" is given a value of +1 for event "making a cake". With the relevancy matrix, we use the following equation for predicting the presence of each event in a video:

$$p = \mathbf{R} \times c, \tag{1}$$

where c is a l-dimensional concept classification score vector for the video, and p is a k-dimensional event prediction score vector. In our experiments, we only consider the case where each video contains a single event, and thus the final selected event corresponds to the index in p with maximum prediction score.

This simple event prediction process helps the recounting by preventing conflicts in concept occurrence caused by noisy classification. For example, the automatic concept classification may predict that the background scene of a video changes from indoor "kitchen" to "outdoor", and then flips back to "kitchen" again. In this case, if the video is predicted with an event "making a cake", the noisy "outdoor" classification will not be used in recounting since it conflicts with "making a cake" according to the event-concept relevancy matrix $\mathbf{R}$.

Next we describe the rules for generating textual descriptions. In this pilot exploratory work, we consider three types of concepts: human action concepts (e.g., "walking"), scene concepts (e.g., "kitchen"), and audio sound concepts (e.g., "cheering"). The templates used for recounting are predefined based on these concepts. Our first template is based on the classification of concepts "crowd" – if it is detected, we use "Several people are..." as subject phrase. Otherwise we use "Someone is...". The subject phrase is then concatenated with the action and scene phrases accordingly based on the identified action/scene concepts. For example, if "crowd" and "walking" concepts are identified simultaneously, then we output "Several people are *walking*". If both action (e.g., "walking") and background scene (e.g., "kitchen") concepts are identified, we form sentences like "someone is *walking* in the *kitchen*". Sometimes scene concepts may be identified but no people related concept is found. In this case, we use phrases describing scene setting only, such as "The background is a *baseball field*". In addition, if there are multiple scene concepts identified, we only generate a sentence based on the one with the highest confidence. This way we can further reduce noisy classification and make the final recounting more concise and less wordy. The overall recounting mecha-

Table 1: Audio-visual concepts used in our experiments.

Human Action Concepts	Scene Concepts	Audio Concepts
walking	kitchen	outdoor rural
running	outdoor with grass/tree	outdoor urban
squatting	baseball field	indoor quiet
standing up	crowd	indoor noisy
making stuff with hands	cake close-up	speech comprehensible
batting baseball		music
		cheering
		clapping

nism is pretty generic — it is very easy to incorporate more concepts and/or templates into the framework for more detailed video recounting.

We use the same set of simple templates described above to generate descriptions at both 10-second clip level and entire video level. The entire video level recounting is generated by removing all the redundant (duplicate) phrases from the 10-second level descriptions.

3. EXPERIMENTS

3.1 Experimental Setup

Experiments are performed on Internet videos from NIST TRECVID 2010 multimedia event detection (MED) task[1]. The MED task was set up to foster automatic content recognition and recounting techniques for complex videos. Three events were defined in MED 2010: "assembling a shelter", "baseball batting a run in", and "making a cake". Our target in this work is to automatically produce textual descriptions for the MED test videos that contain at least one of the events (140 videos in total). For the audio-visual concepts, we use 19 concepts defined by Columbia University [2]. Labels of these concepts are also provided by [2] on 7,156 10-second clips from 565 training videos, i.e., each 10-second clip of the 565 training videos has been manually labeled w.r.t. the 19 concepts[2]. We use STIP feature for human action concept classification, SIFT feature for scene classification, and MFCC feature for classifying audio concepts. The classifiers are then applied to 10-second clips of the 140 test videos to generate audio-visual concept classification scores. During recounting, a concept is treated as true if it has a classification score larger than 0.4.

We conduct subjective user studies to evaluate the recounting performance. Figure 2 displays the user-interface design. Recounting score ranges from 1 to 5 (from very bad to very good), indicated by the number of color-highlighted hearts shown in the figure. Before the evaluation, the users were instructed and guided about the score assignment in detail using several example videos. During the evaluation, videos were randomly assigned to users and each video was scored by at least three independent users. As introduced earlier, evaluation is conducted on both 10-second clip level and entire video level. The evaluation interfaces for both are similar, except that for the 10-second level recounting, the textual descriptions below the video are online updated as the video is playing.

3.2 Results and Analysis

We first report event prediction performance based on the audio-visual concept classification outputs, as described in Section 2.2. Since concept classification is performed on 10-second clip level, there are two ways to consolidate the predictions for generating the overall concept vector, c, by using the "mean" and the "max" values of concept scores over all 10-second clips in a video. We report performance of both strategies. Overall, the event prediction accuracy of both "mean" and "max" strategies is equal to 89.3%. Table 2 summarizes the number of misses and false alarms for each event. "Max" has less misses for event "batting a run in". This is because some concepts only occur once throughout the baseball videos, e.g., "running" or "cheering". On the other hand, the "max" strategy gives negative impact for event "making a cake". This may be due to the fact that "making a cake" is a long-term event with many concepts happening continuously (e.g., "kitchen", "making stuff with hands", etc.), for which "mean" strategy is more reliable. Since there is no clear winner between the two, we only report recounting results using "max" in the following.

Now we summarize the recounting evaluation results from user studies. In total there are 43 human evaluators involved in this subjective scoring process. Figure 3 shows the user scores for both 10-second clip level and entire video level recounting. Overall, the results are very encouraging – about 1/3 of the videos received full score 5 hearts, and another 1/2 of them were given 3–4 hearts. Comparing videos of the three events separately, "batting a run in" videos have more positive feedbacks than the other two, which is probably due to the fact that this event has relatively more consistent (and easy-to-identify) background scene pattern. Videos with the "assembling a shelter" event, on the other hand, are more difficult to recount since the scene/object settings of this event have very large intra-class variations.

Figure 2: User evaluation interface. A video is displayed on the top and its textual description is given under it. Users were asked to score the recounting quality by assigning 1-5 hearts shown at the bottom.

Table 2: Concept-based event prediction results. We report the number of misses and false alarms (FA) for each event, using two consolidation strategies: "mean" and "max".

Event (true positive video #)	"Mean"		"Max"	
	#Miss	#FA	#Miss	#FA
Assembling a shelter (46)	2	13	2	13
Batting a run in (47)	6	1	2	2
Making a cake (47)	7	1	11	0

[1]http://www.nist.gov/itl/iad/mig/med10.cfm

[2]The audio-visual concept annotations were downloaded from http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/TRECVID2010MED/TRECVID2010MED.htm

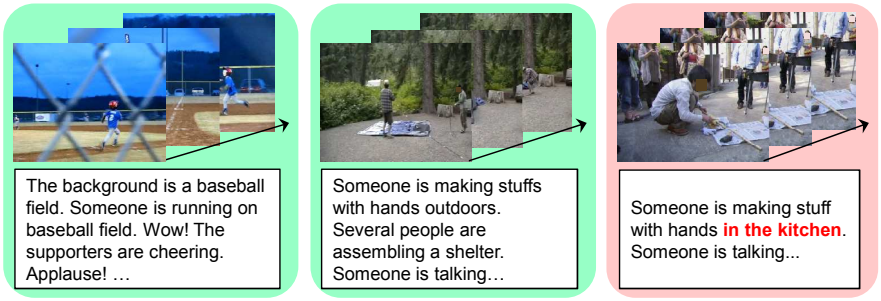
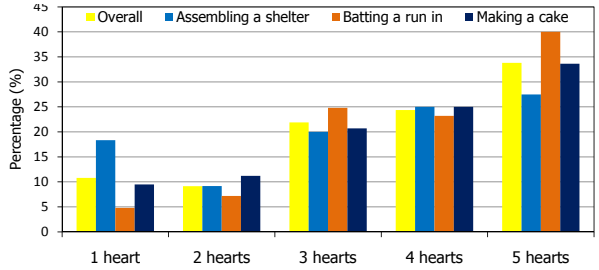
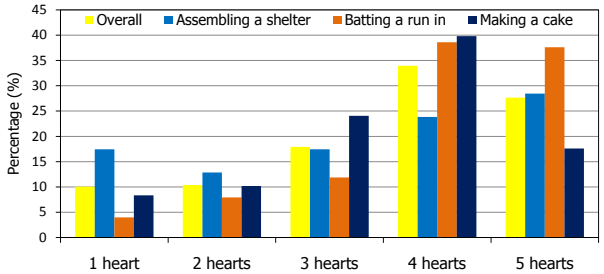


Figure 4: Recounting result examples. The left two are good, while the right one is a bad result example, where the outdoor street scene was mistakenly recognized as indoor kitchen. This is mainly due to the low coverage of our current audio-visual concept library which only contains a specific outdoor concept "outdoor with grass/tree".



(a) 10-second clip level recounting



(b) entire video level recounting

Figure 3: User evaluation results of automatic content recounting on (a) 10-second clip level, and (b) entire video level. Scores were marked by 1-5 hearts indicating recounting quality varying from very bad to very good. Overall, users involved in the evaluation were quite satisfied with the results. Entire video level recounting is a bit worse than that on the 10-second level since it is more challenging to generate short and descriptive overall summaries.

User evaluation results of the two levels of recounting do not differ too much, except that the entire video level recounting received a smaller number of full scores (5 hearts). This indicates that there may be some missing clues or conflicts resulted from the generation of the overall recounting, which deserves deeper analysis that will be done in our future work. Figure 4 shows some examples with both good and bad recounting outputs.

4. CONCLUSIONS

We have conducted a pilot study towards the challenging goal of generating video content descriptions automatically. Our focus is particularly on the utilization of audio-visual concept classifiers obtained based on the state-of-the-art content recognition techniques. Using a set of Internet videos and 19 audio-visual concept classifiers as a showcase, we have observed that simple rule-based recounting method is already able to generate very encouraging results. Our main contribution of this work is to show that the imperfect classification of audio-visual concepts can be exploited for effectively describing complex video contents. There are still many ways to improve the current prototype system. Significant ones include the utilization of advanced natural language processing (NLP) techniques for generating more elegant textual descriptions, and the extension of our concept library to cover more real-world audio-visual settings.

Acknowledgement

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709). YGJ is supported by Fudan University New Faculty Start-up Grant.

5. REFERENCES

[1] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9:958–966, 2007.

[2] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[3] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50:171–184, 2002.

[4] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, 2005.

[5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[6] K. Lee and D. P. W. Ellis. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech and Language Processing*, 18:1406–1416, 2010.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision*, 60:91–110, 2004.

[8] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[9] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.