Detecting Complex Events in User-Generated Video Using Concept Classifiers

Jinlin Guo, David Scott, Frank Hopfgartner, Cathal Gurrin
CLARITY and School of Computing,
Dublin City University,
Glasnevin, Dublin 9, Dublin, Ireland
{jinlin.guo, david.scott, frank.hopfgartner, cathal.gurrin}@computing.dcu.ie

Abstract

Automatic detection of complex events in user-generated videos (UGV) is a challenging task due to its new characteristics differing from broadcast video. In this work, we firstly summarize the new characteristics of UGV, and then explore how to utilize concept classifiers to recognize complex events in UGV content. The method starts from manually selecting a variety of relevant concepts, followed by constructing classifiers for these concepts. Finally, complex event detectors are learned by using the concatenated probabilistic scores of these concept classifiers as features. Further, we also compare three different fusion operations of probabilistic scores, namely Maximum, Average and Minimum fusion. Experimental results suggest that our method provides promising results. It also shows that Maximum fusion tends to give better performance for most complex events.

1. Introduction

Due to the increasing popularity of Web 2.0 applications such as YouTube[1] and Flickr[2] that allow users to instantly upload their own multimedia content, the amount of user generated videos (UGV) on the web has been increasing significantly. Unfortunately though, most user generated video content is rather unstructured and poorly annotated, rendering access and retrieval of this content a challenging task. A promising approach to address this problem is to automatically create annotations, e.g., by describing the events that are shown in the videos. Event detection is particularly important to understand the semantics of video content, which is helpful for video summarization, indexing and retrieval purposes.

Most of the existing work on event detection either focus on simple events such as *dancing* or is limited to specific types of video, such as movies, sports or news video with rich domain-specific knowledge for constructing event models. Detecting simple events in video data is usually considered to be a semantic concept detection task, and therefore, generic semantic concept detection techniques are adopted [3, 11, 13].

Most approaches to detect events are domain dependent, i.e., they rely on specific domain knowledge and textual metadata information such as movie scripts or closed caption. Since complex events in these types of video can be regarded as a spatial, temporal, and logical interaction of multiple objects, many techniques focus on tracking the objects and analyzing their activity patterns. Common techniques to identify these complex event patterns are transition based approaches such as Hidden Markov Models (HMM) [15], Petri-nets [1] and Bayesian Networks [14].

Specifically, for event detection in movies, recent work [6, 2] incorporates visual information, closed-caption transcripts, and movie scripts to automatically annotate and classify movie video. For event detection in sports video, most of the previous work rely on domain knowledge. In [8], Poppe et al. summarize existing work and present a generic architecture for automatic annotation of broadcast sports video. There is also a lot of work for event analysis in news video [16, 17]. By exploiting the available audio, visual and closed-caption cues, the semantically meaningful highlights in a news video are located and event boundaries are extracted. In [1], Bai et al. utilize an audio-visual feature-based framework for event detection in broadcast video of multiple different field-based sports. The evidence gathered by the feature detectors is combined by means of a Petri-Net, which infers the occurrence of an event. In their approach, however, they heavily rely on sports specific domain knowledge.

In [18], Zhang et al. present a generic event detection approach based on semi-supervised learning, which can be applied to sports, movie and news video. Their method jointly explores small-scale expert labeled videos and large-

[1] http://www.youtube.com
[2] http://www.flickr.com

scale unlabeled videos to train the models to detect video events. The labeled videos are obtained from the analysis and alignment of well-structured video related text (e.g. movie scripts, web-casting text). Unlabeled data are obtained by querying related events from the video search engine (e.g. YouTube) in order to give more distributive information for event modeling.

While above approaches provide decent results to detect simple events, they are doomed to fail when they are applied to identify more complex events, i.e., events that a) involve several **objects** (or persons) with loosely or tightly-organized complex **activity**, b) generally occur under specific **scene** settings, and c) evolve in different patterns over time. In 2010, TRECVid [7] started a new task named Multimedia Event Detection (MED)[3], which aimed to fostering automatic complex event detection in Internet video. Now, the task has attracted a lot of participants, and different techniques and mixed performance were reported, such as [4, 9]. The video data used for this task was collected by the Linguistic Data Consortium[4]. It consists of publicly available, user-generated content posted to various Internet video hosting sites.

In comparison to produced video such as news and sports video, specific characteristics existing in UGV also make it more difficult to detect video events. UGV is usually:

- **Of lower quality**: Due to the uncontrolled capturing conditions and different capture devices, UGV are most of the time of lower quality than professionally produced video. For example, UGV suffer from irregular camera motion and fuzzy backgrounds.

- **Less structured**: UGV are not as well structured as news and sports videos. News and sports video is produced after careful processing, and the structure is clear (see Fig. 1). However, UGV is usually captured using personal video recorders by different users, and there is most of the time no post-production before uploading to the video sharing websites. Therefore, there is less structural information existing in the UGV and it is more difficult for methods based on state transition learning the patterns.

- **More diversified**: The diversity of UGV includes internal diversity and external diversity. Internal diversity means video reflecting the same topic can be completely different (see Fig. 2). Video is a tool that can be used to express ideas [10]. While producing video such as soccer video, common editing ideas (or rules) have been established that the producers (experts with rich experience in recording and editing video) will comply with. However, UGV can be recorded under uncontrolled capturing conditions by a diverse group of producers (e.g., website users without much knowledge in creating and managing video). External diversity comes from the diversity of topics. Because of the open and sharing of the video websites and the diversity of authors, video about nearly any topic can be uploaded. That is, the UGV can be about anything, and anyone can be a star, from lip-synching teenage girls to skateboarding dogs.

- **not described following domain-specific rules**: There is far less domain-specific knowledge (DSK) for analyzing and processing UGV. As described in Section 1, the most successful event detection methods utilize DSK within the video to detect events.

- **annotated with less metadata**: In the scenario of UGV, the metadata information is not always available. For examples, speech transcripts are absent in UGV, and instead, user tags, which could be noisy and redundant, are available.

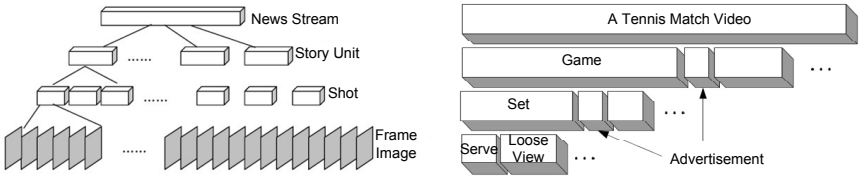
Furthermore, much more UGV data than produced video can be found on the Web. According to the official statistics from YouTube, 48 hours of video are uploaded every minute by their users, resulting in nearly eight years of content uploaded every day. Within one month, more video is uploaded to YouTube than the three major US networks created within the past 60 years[5].

Due to the new features mentioned above, it is difficult to detect complex events in UGV. Our approach is to explore complex event detection in UGV using concepts which are highly relevant to the complex events. The rational can be explained by the example of the event *Getting a vehicle unstuck* as illustrated in Fig. 2. This event can happen in different *outdoor* locations, either with or without *person(s) pushing* or *pulling* a vehicle. Despite these differing features that describe the event, we can model it by identifying certain highly relevant concepts such as *car* or *outdoor*. Hence, concept detectors can be used to identify aspects of an event. Research in automatic detection of semantic concepts has now reached the point where hundreds of concept detectors can be obtained in generic fashions, albeit with mixed performance, which offer novel opportunities for video retrieval.

Within this work, we first manually select a series of concepts including scene, object and human action that are suitable to depict the essence of certain events and construct classifiers for these concepts. Finally, complex event detectors are learned by using the concatenated probabilistic output scores of these concept classifier as features. Moreover, we compare three different fusion operations of probabilistic output scores of concept detectors, namely *Maximum*,

[3] http://www.nist.gov/itl/iad/mig/med11.cfm
[4] http://www.ldc.upenn.edu/About/

[5] http://www.youtube.com/t/press_statistics, accessed on 30 April 2012



(a) A structural method for broadcasted news video (b) A structural method for a tennis match video

Figure 1. Two structural methods for broadcasted news video and tennis match video respectively

Figure 2. Visual diversity of the same video event: Getting a vehicle unstuck

Average and *Minimum* fusion. In the remainder of this paper, we provide the details of our approach, followed by the experiments and discussions.

2 Modeling Complex Event in Semantic Concept-Space

Fig. 3 depicts the flowchart of our event detection technique. In this chart, we divide the process of complex event detection into three steps: selection of relevant concepts (including object, scene and human action), concept learning, and event detection with concept scores combination.

2.1 Selection of highly relevant semantic concepts

While in the approach of event detection using concepts, there is an issue of relevant concept selection, we will not elaborate this challenging question due to space limitation. In this work, we manually select relevant concepts for given events since knowledge about the world is implicit in human minds and retrieval systems can exploit this knowledge by asking humans to select appropriate concepts for complex event queries. Specific steps for concepts selection are as follows. We first manually select twenty positive UGV for Event e from the training set. Each video is divided into multiple 6-sec subclips for human action learning, and we extract one keyframe every four seconds for other concepts learning. Then, two users with rich experience in multimedia retrieval are asked to inspect these keyframes and video subclips for each positive video, and to give a list of concepts which they find highly relevant for Event e. The object and scene concepts are from viewing the statistic keyframe images, whilst the human action concepts are from the video subclips. This is similar to generating document vectors in text retrieval field to find the most discriminative words. Therefore, for Event e and the selected positive UGV, a concept-event matrix is obtained by the same method to produce the *tf-idf* (term frequency-inverse document frequency) matrix.

Since this concept-event matrix is generated in a collection in which all the videos share the same topic (Event e), the weights in the matrix measure the importance of the concepts to the topic. Based on the weights in the concept-event matrix, we select the top n most important concepts for Event e.

2.2 Visual Concepts and human actions learning

We develop automatic classifiers for identifying these concepts selected in Section 2.1 for all events following state-of-the-art approaches for semantic concept detection in [6, 11]. In this work, we only consider visual features, it can be easily extended to audio concepts though. Specifically, we extract two kinds of feature descriptors: static OpponentSIFT [12], 3D spatial-temporal interest points (STIPs) [5] for object/scene concepts and human action concepts learning respectively. The OpponentSIFT feature is an extension of the Scale-Invariant-Feature Transform (SIFT) feature to the opponent color space and is known to be a good performing single feature for concept detection [12]. The STIP feature effectively captures space-time volumes where the image values have significant local variations 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. We use concatenated HOG and HOF feature (162 dimensions) as the final descriptor for each STIP.

After extracting the OpponentSIFT and STIP descriptors, the popular bag-of-visual-word (BoVW) representation is then applied to convert the two sets of descriptors separately into two fixed-dimensional feature vectors. Hi-

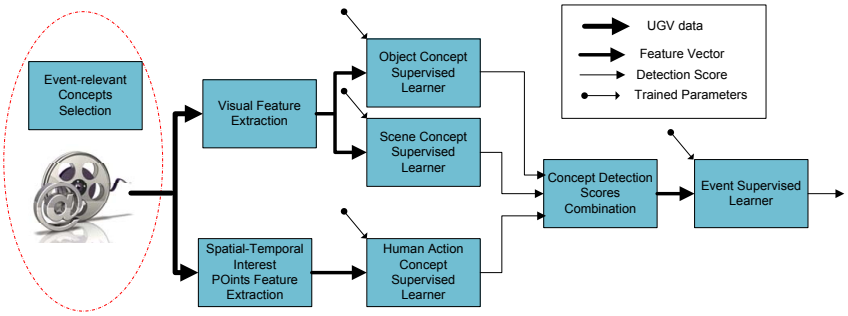


Figure 3. The flowchart of complex event detection in UGV based on concept classifiers.

erarchical K-means generate a visual vocabulary of 4096 words for STIP features and a smaller visual vocabulary of 512 visual words OpponentSIFT respectively.

Since processing all video frames will be computationally very expensive, we sample one keyframe every four seconds, and extract OpponentSIFTs descriptors on these keyframes using 2 spatial pyramids of 2×2 and 1×3. Therefore, a 3584-dimension feature vector is generated for each keyframe ($2 \times 2 \times 512 + 1 \times 3 \times 512$) after the vector quantization process [12]. Each video is divided into multiple 6-second clips and STIP descriptors are extracted.

In the construction of BoVW representation, we use soft-assignment [3] to assign each feature descriptor to four nearest visual words with different weights. Annotations of training samples for object and scene concepts are performed on the keyframe images, whilst annotations of human action concepts are performed on 6-second video subclips. With these BoVW feature representation and annotations, we use SVM classifier for concept learning. The χ^2 kernel is chosen for SVM for its good classification results. The concepts used in our experiments and their training data will be introduced in the experiment section.

2.3 Concept Detection Scores Combination

For each video clip, we have multiple probabilistic output scores for each concept, where each score corresponds to one keyframe or video subclip in this video clip. We use three common fusion strategies to generate the final detection score for one concept, namely, *Maximum*, *Average* and *Minimum*. Given a video clip u, the final probabilistic output score for concept c is

$$p^*_{u,c} = FusionOperator(p^1_{u,c}, p^2_{u,c}, ..., p^l_{u,c})$$

where *FusionOperator* can be *Maximum*, *Average* or *Minimum* operator. $p^i_{u,c}$ is the probabilistic output score of the classifier of concept c on the i^{th} keyframe image (or video subclip) in the video clip u. l is the number of keyframe images (or video subclips) in the video clip u.

Table 1. Event names and the number of positive samples in the training set

Event Name	Positive Samples #
Birthday party	231
Changing a vehicle tire	121
Flash mob gathering	192
Getting a vehicle unstuck	151
Grooming an animal	143
Making a sandwich	187
Parade	192
Parkour	136
Repairing an appliance	146
Working on a sewing project	133

For fusion operator f, we concatenate the final probabilistic output score of each concept, into vector v^f_u describing video clip u. Then v^f_u serves as the input for a χ^2-kernel SVM, which trains event classifiers. Therefore, for each event, we have three classifiers corresponding to *Maximum*, *Average* or *Minimum* fusion of probabilistic output scores respectively. A comparison of event detection performance is presented in the next section.

3 Experiments

3.1 Experiment Setup

We evaluate the framework using Internet videos from the NIST TRECVid 2011 MED task. In Table 3, we list the ten complex events that we evaluated and the number of positive video clips in the training set for each event. The testing set consists of 32061 video clips. Using the method described in Section 2.1, we manually select 50 concepts for event detection as listed in Table 2.

Training a large number of concept detectors requires a lot of manual annotation work, and the key problem is that the positive samples are usually very rare in the training set.

According to the relationship between concepts and events, we can easily obtain the positive samples for concept classifiers by only annotating the positive samples for events, hence, speeding up the annotation. After concatenating the probabilistic output score of each concept, a second-stage annotation is performed on the event level.

For measuring detection performance, we use the common measure in the multimedia retrieval field, average precision (AP), which summarizes the recall-precision curve.

3.2 Results and Analysis

We first report and compare the prediction performance of the ten most complex events. As described in Fig. 4, the two events *Birthday party* and *Changing a vehicle tire* achieve the highest average precision of 0.0898 and 0.098, respectively, using *Maximum* fusion. We conclude that for these two events, the relevant object concepts, (namely, *person*, *streamers*, *balloons*, *birthday cake* for event *Birthday party*, and *vehicle*, *tire*, *lug wrench*, *hand*, *screwdriver* for *Changing a vehicle tire*) capture the characteristics of these two events very well. And the classification performance for these concepts gains a lot because of our spatial pyramid segmentation (see Section 2.2) of the keyframe images. For the event *Flash mob gathering*, the three fusion operations report the worst performance. In this event, a large group of people assemble suddenly, perform a group dance and then disperse quickly. However, in our approach, no spatial/temporal partitioning is used since the STIP feature does not capture the temporal feature of this event. Moreover, it is difficult to tell the difference between this event and other scene such as *walking crowd towards a certain direction*. Therefore, many false alarms are reported. The performance for the event *Parade* achieves the third position. We deem that this is the case because the STIP feature describes this event well and there are more positive video clips in the testing set (see Table 3). Overall, for most of the events, the APs range from 0.03 to 0.06. This is mainly caused by the quality of the concept classification performance.

As shown in Fig. 4, we also see that the *Maximum* fusion outperforms the other two fusion operations for seven of the ten events. In the three cases where *Maximum* fusion is outperformed by *Average* and *Minimum* fusion, the performance difference is not significant. In a video clip about a complex event, the relevant concept does not appear in every keyframe image or video subclip. Therefore, *Maximum* fusion reaches better performance than *Average* and *Minimum* fusion.

As displayed in Table 3, we also list the number of true positives at the Top 10 and Top 100 of the ranked predictive results, as well as the number of the true positive video clips in the testing set. For the events *Flash mob gathering* and *Grooming an animal*, our approach did not retrieve any results. This may be explained by the fact that no human action concept is selected for this event; only the objects *hand*, *animal* and scene object *indoor* are used to characterize these events. Consistent with the results shown in Fig. 4, there are four true positives in the Top 10 ranked results for event *Birthday party* and *Changing a vehicle tire* respectively. For the event *Making a sandwich* and *Working on a sewing project*, three relevant video clips are in the Top 10 ranked results respectively. These results suggest that this method can still retain the potential in generating promising performance using concept classifiers. However, more work should be attached on increasing the performance of concept classifiers.

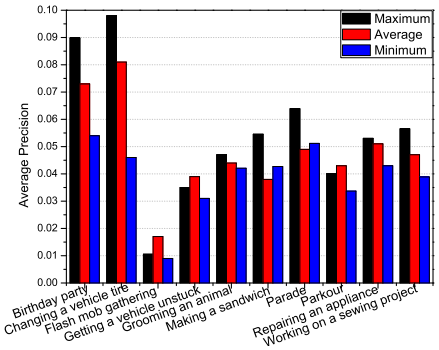


Figure 4. Prediction performance comparison of ten complex events for three fusion operations.

Table 3.

Event(true positive video #)	Top 10	Top 100
Birthday party(186)	4	23
Changing a vehicle tire(111)	4	24
Flash mob gathering(132)	0	6
Getting a vehicle unstuck(95)	1	10
Grooming an animal(87)	0	10
Making a sandwich(140)	3	18
Parade(231)	2	20
Parkour(104)	1	11
Repairing an appliance(78)	2	13
Working on a sewing project(81)	3	13

4 Conclusion and Future Work

In this paper, we summarize the new characteristics of UGV and explore how to utilize the concept classifiers to detect complex event in UGV. Further, we also compare three different fusion operations of probabilistic out-

Table 2. Concepts selected for complex event detection

Object	streamers, person, road, vehicle, tire, lug wrench, animal, hand, snow, sink, bread, plates, jack, buildings, walls, stairs, sewing machine, balloons, birthday cake, screwdriver, table
Scene	indoor, outdoor with trees or grass visible, outdoor with cityscape, crowd, street, waterfront
Human Action	pushing, pulling, digging, slicing, spreading condiments on bread, jumping, clapping, unscrewing screws, rolling, bending over, dancing, eating, kneeling, playing games, running, turning lugwrench, unscrewing bolts, walking, sewing, singing, holding objects, cutting, pressing

put scores of concept detector, namely *Maximum*, *Average* and *Minimum* fusion. Experimental results on a large-scale dataset show that this method retains the potential in generating promising performance. It also shows that *Maximum* fusion tends to give better performance for most complex events. There are many issues to be addressed in the future in using the concept classifier approach for modeling and detecting the complex events in UGV. In future work, we will focus on the following aspects. The first is automatic selection of relevant concepts. Moreover, we will extend our approach by incorporating audio concepts. Another plan is to increase the performance of the concept classifiers by incorporating multiple low-level features and different training sets.

Acknowledgement

Thanks to the Information Access Disruptions (iAD) Project (Norwegian Research Council), Science Foundation Ireland under grant 07/CE/I1147 and the China Scholarship Council for funding. Moreover, many thanks to the HMA group[6] for their help in the annotation effort.

References

[1] L. Bai, S. Lao, A. F. Smeaton, N. E. O'Connor, D. A. Sadlier, and D. Sinclair. Semantic analysis of field sports video using a petri-net of audio-visual concepts. *Comput. J.*, 52(7):808–823, 2009.

[2] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV (4)*, pages 158–171, 2008.

[3] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[4] 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.

[5] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[7] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quenot. TRECVID 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[8] C. Poppe, S. De Bruyne, and R. Van de Walle. Generic architecture for event detection in broadcast sports video. In *Proceedings of the 3rd international workshop on Automated information extraction in media production*, AIEMPro '10, pages 51–56, New York, NY, USA, 2010. ACM.

[9] D. Scott, J. Guo, C. Foley, F. Hopfgartner, C. Gurrin, and A. F. Smeaton. TRECVID 2011 experiments at Dublin City University. In *NIST TRECVID Workshop*, 2011.

[10] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl.*, 25(1):5–35, 2005.

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

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, 2010.

[13] J. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Comparing compact codebooks for visual categorization. *Computer Vision and Image Understanding*, 114(4):450–462, 2010.

[14] F. Wang, Y.-F. Ma, H. Zhang, and J.-T. Li. A generic framework for semantic sports video analysis using dynamic bayesian networks. In *MMM*, pages 115–122, 2005.

[15] L. Xie, P. Xu, S. fu Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden markov models, 2003.

[16] Y.-X. Xie, X.-D. Luan, S. Lao, L.-D. Wu, P. Xiao, and Z.-G. Han. A news video mining method based on statistical analysis and visualization. In *CIVR*, pages 115–122, 2004.

[17] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1985–1997, 2008.

[18] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu. A generic framework for event detection in various video domains. In *ACM Multimedia*, pages 103–112, 2010.

[6] `http://hma.dcu.ie/HMA/Home.html`