

# Automatic Event Detection in User-Generated Video Content: A Survey



Alamuru Susmitha, Sanjay Jain, and Mihir Narayan Mohnaty

**Abstract** The aim of event detection is to identify interested events in a user-generated content using multiple modalities automatically. However, it is a challenging task particularly when videos are captured in a restricted environment by nonprofessionals. Such videos suffer from poor quality, deprived lighting, blurring, complex camera motion chaotic background clutter, and obstructions. However, with the rise of social media, there is rising popularity of user-generated videos on the Web day-by-day. Each minute, 300 hours of user-generated video are uploaded on you tube due to which people find difficult to search the appropriate content among a large number of videos. Therefore, solutions to this problem are in great demands. In this paper, we study existing technologies for event detection in user-generated videos using multiple modalities. This paper provides key points about feature representations across different modalities, classification techniques.

**Keywords** Event detection · User-generated video content · Modalities · Video indexing · Video retrieval · Video summarization

## 1 Introduction

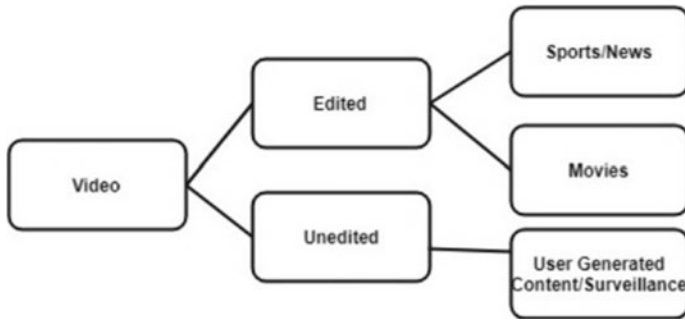
The growth of social media over the last decade attracted users to create and to immediately post their own content with no restrictions on the content. As a result, the user-generated content has been increasing rapidly on the Web.

---

A. Susmitha · S. Jain (✉)  
CMR Institute of Technology, Bengaluru, India  
e-mail: [principal@cmrit.ac.in](mailto:principal@cmrit.ac.in)

A. Susmitha  
e-mail: [susmitha.academic@gmail.com](mailto:susmitha.academic@gmail.com)

M. N. Mohnaty (✉)  
ITER, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India  
e-mail: [mihir.n.mohanty@gmail.com](mailto:mihir.n.mohanty@gmail.com)



**Fig. 1** Overview of types of videos

The most widely used user-generated contents are text messages (tweets on Twitter), audio–speech and music, images or pictures and short sequences of moving images also known as video clips. User-generated content is valuable resources of information capturing people’s interests, thoughts, and actions. Automatic video understanding [1] is crucial among them. Different types of videos [2] can be produced as shown in Fig. 1.

## 2 Video

Video is a short sequence of moving images and audio. A video is the only asynchronous arrangement of several frames, each frame being a 2D representation. So, the important unit in a video is a frame. The video can be considered as a gathering of numerous scenes as shown in Fig. 2, [3] where the scene is an accumulation of shots that have a similar setting. Therefore, video consists of an enormous amount of content in terms of scenes, shots, and frames.

### 2.1 *User-Generated Video Opportunities and Challenges*

In general, user-generated video is of poor quality and less organized. As there are restricted capturing situations, they may be of poor quality than professionally edited videos. Both sports and news videos are made after proper editing. However, most UGV is usually captured using own smart phones by individual users, and without any editing, they will be uploaded on the web. So, UGV is unstructured.

According to Twitter statistics, each minute, twitter dynamic users create about 500 million tweets every day and YouTube users post 300 h of videos due to which people find it difficult to look for the appropriate content among a huge number of videos. Therefore, the need for automatic event detection in user-generated video

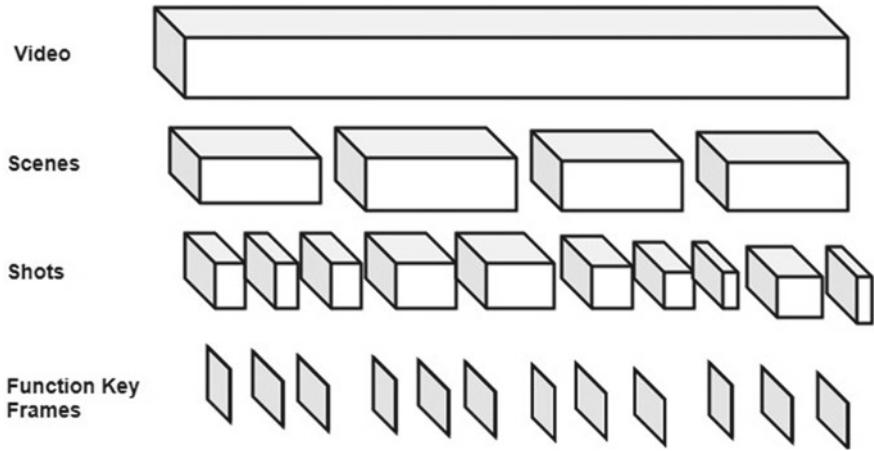


Fig. 2 Internal complex architecture of video

data is obvious in many computer vision applications. Finding a solution to automatically figure out the events captured in this large collection of videos is not an easy task. Therefore, user-generated video data provides both opportunities as well as challenges. The primary challenge is how to handle such huge data in a proficient manner. In addition to that, it is complex to search for videos based on user interested specific events.

### 3 Event Detection

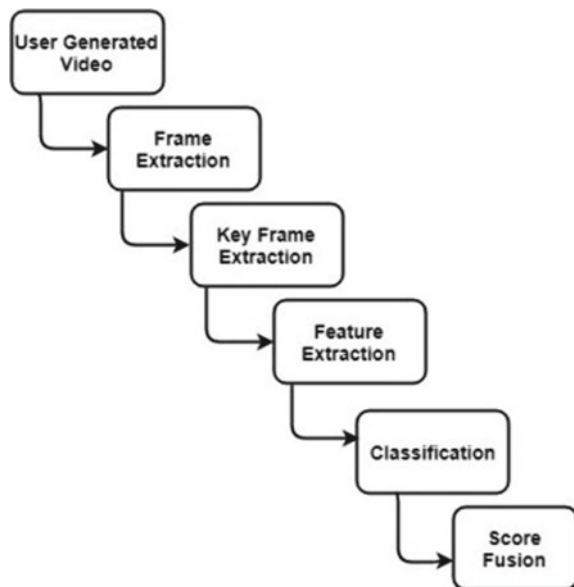
The aim of event detection in user-generated videos is to automatically detect the video clips based on user interested specific events in a given video. Event detection using single modality features or fusion of multiple single modality features can be performed in sports, news, movies, surveillance, large-scale videos, and user-generated videos. User-specific interested events detection in UGV is a quite challenging task because the videos are captured in a restricted environment by non-professional users. They are unstructured and unconstrained. They suffer from poor quality, deprived lighting, blurring, complex camera motion, chaotic background clutter, and obstructions. An event is a significant occurrence specified by the user. It happens at a certain place at a certain time. For example, human actions like jumping, running, and object-related events like kicking a ball, riding a horse, etc. Events are characterized by its type, time, location, and description. The process of identifying the occurrence of an event in a natural or manmade content is called event detection. Detecting substantial change in sea-level is a natural event detection and detecting events like smiling, frowning, etc., in images or videos is called manmade event detection. The social media platforms allow millions of people to use them

daily to communicate and share information ranging from world level information, for example, the World Cup, to personal information like wedding, graduation. A massive amount of data is created by individual users in the form of texts, videos, and photos. The research done so far could give efficient solutions for large data storage but retrieving, handling, and processing of such a large amount of data particularly in videos are still a challenge. So monitoring and evaluating the user-generated video content can produce undoubtedly valuable information.

### 3.1 General Event Detection Methodology

- Frame extraction—Depending upon the size of the video, it consists of a number of frames. Frame extraction is to represent the video in terms of image frames (Fig. 3).
- Key frame extraction—Extraction of key frames is the basic step in video-related tasks to get rid of the duplicate frames with unnecessary data. The extracted key frames represent the characteristics of the video.
- Feature extraction—It extracts important features from video data to enable semantic understanding. Using feature extraction process, visual, audio, and audio-visual features can be extracted. Spatial, transform, color, texture, shape, edge and boundary, structure, layout, and motion are some of the visual features. The most common audio classes in videos are speech, silence, music and the





**Fig. 3** General event detection methodology



combination of later three. The audio features can be embedded with low-level visual features for key frame extraction.

- **Classification**—Classification is done using classifiers. After extracting features, classifiers generate scores based on different model formulations and set of features. These are the most significant concepts in image processing; these are computer-based mathematical algorithms developed to encounter the required performance level, at trained data set with a given amount of time. The classifier is trained in such a way that the system must easily differentiate the datasets.
- **Score fusion**—It combines scores computed from different features from different modalities. Based on this, decision can be made.

**Table 1** Overview of different modalities

S.No.	Modality	Example	Description and related work
1	Text		Most of the research work done so far on the single text modality conveyed that the textual features from a video were obtained using either automatic speech recognizer (ASR) or optical character recognition (OCR) [5]
2	Audio (speech and music)		Mel-frequency cepstral coefficients (MFCC) is a popular and standard feature of audio [6]
3	Visual (image)		The most widely used low-level visual features are spatial-temporal interest points (STIP) [7], scale-invariant feature transform (SIFT) [8], histogram of gradients (HOG), histogram of optic flow (HOF) [9], color, GIST, independent subspace analysis (ISA), geometry texton histogram (GTH), transformed color histogram (TCH), local binary patterns (LBP), and speeded up robust features (SURF) [10]
4	Motion		It is the representation of kinetic energy. It is used to measure the variation of pixels within a shot, direction of the motion, and histogram magnitude

To improve the accuracy of event detection in user-generated videos, features can be extracted from four different modalities: text, audio, visual, and motion. Overview of different modalities with an example and description is given in Table 1 [4].

## 4 Multimodal Event Detection

Multimodal event detection is based on multimodal fusion techniques, i.e., a fusion of multiple features from different modalities which are referred to as multimodal fusion. The fusion of various modalities may give related important information and, therefore, it is better to know which modalities will contribute a major role for accomplishing an analysis task. The fusion of both visual and audio features along with the video textual data in a user-generated video will improve the accuracy of event detection. Hence, the extraction of helpful features from video one by one to get better recognition of events is a further important task due to the discrete features of the concerned modalities.

As different modalities possess different characteristics, it is better to consider the confidence levels of the modalities in completing the required multimedia applications. The multimodal event detection framework [11] is shown in Fig. 4.

In Table 2, we listed the papers that used visual features alone and features combined from multiple modalities for different multimedia tasks.

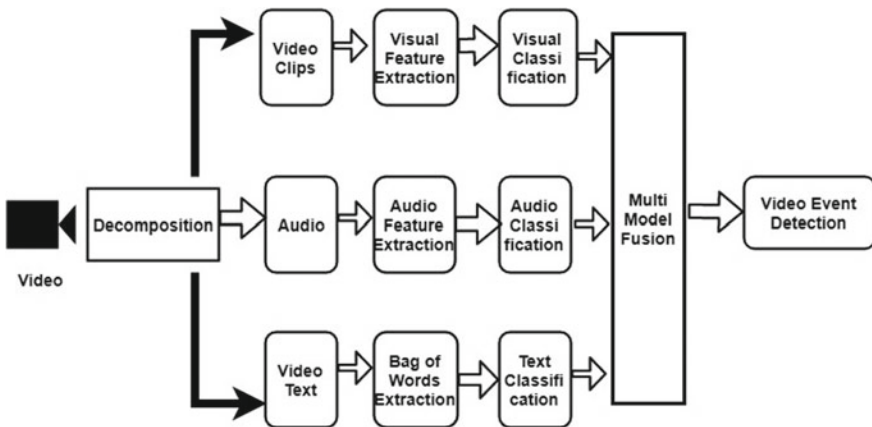


Fig. 4 Framework of multimodal event detection

**Table 2** Overview of related work

S.No.	Paper	Modalities used	Classifier	Multimedia task
1	[7]	Low-level video (spatiotemporal descriptors) and audio (Mel-frequency cepstral coefficients) features	CRF-based discriminative learning	Event detection in sports video
2	[12]	Visual features (HOG, CEDD, color histogram, texture, and wavelet)	Decision tree (DT), multiple correspondence analysis (MCA), and support vector machine (SVM)	Automatic video event detection in disaster data set
3	[9]	Visual features (static opponent scale-invariant-feature transform (SIFT) 3D spatial-temporal interest points (STIPs))	Support vector machine (SVM)	Complex event detection in user-generated video
6	[13]	Video (color, structure, and shape) Audio (MFCC), textual cues	Support vector machine (SVM)	Semantic concept detection
7	[14]	Audio (ZCR, LPC, and LFCC) Video (blob location and area)	Bayesian Inference	Event detection for surveillance
8	[15]	Visual (color SIFT) Audio (MFCCs) (acoustic segment model)	Latent support vector machine (LSVM)	Multimedia event detection
9	[16]	Sensor data modalities (auxiliary sensors)	Support vector machine (SVM)	Interesting event detection in UGV and extracting appropriate information about the recording activity.
10	[17]	Video and audio	Modern convolution neural network (CNN)	Audio-visual salient event detection
11	[18]	Visual, audio, and motion features	SVM with Gaussian kernel	Robust event recognition in videos

## 5 Conclusion

Video is a rich source of information and topics on video data offer a broad range of research applications such as multimodal event detection which further helps in video browsing, video indexing, video summarization, and content-based video retrieval applications. In recent years, multimodal event detection has been receiving widespread research attention because of the exponential increase in the volume of Web video data. This paper focused on multimodal event detection in user-generated

video content where the objective is to detect video clips by the key event happening in the clip by the fusion of different features from different modalities. We discussed a number of challenges that need to be addressed due to the exponential growth of unstructured Web user-generated video content. We have given an outline of the event detection and the general multimodal event detection framework. Key points about modalities, features, classifiers, and fusion techniques were presented. We think that this paper can give important insights for researchers who are just starting to investigate this area.

## References

1. Lavee, G., Rivlin, E., & Rudzsky, M. (2009). Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in the video. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(5), 489–504.
2. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., & Serra, G. (2011). Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*.
3. Rajendra, S. P., & Keshaveni, N. (2014). A survey of automatic video summarization techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 3(1).
4. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*.
5. Brezeale, D., & Cook, D. J. (2007). Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics*.
6. Atrey, P. K., Anwar Hossain, M., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*.
7. Hassan, E., Gopal, M., Chaudhury, S., & Garg, V. (2011) A hybrid framework for event detection using multi-modal features. In *IEEE International Conference on Computer Vision Workshops*.
8. van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9).
9. Guo, J., Scott, D., Hopfgartner, F., & Gurrin, C. (2012). Detecting complex events in user-generated video using concept classifiers. In *CBMI*.
10. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). SURF: Speeded up robust features. In *Computer Vision and Image Understanding*.
11. Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., & Shah, M. (2013). High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*.
12. Pouyanfar, S., & Chen, S.-C. Automatic video event detection for imbalance data using enhanced ensemble deep learning.
13. Adams, W., Iyengar, G., Lin, C., Naphade, M., Neti, C., Nock, H., & Smith, J. (2003). Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP Journal on Advances in Signal Processing*.
14. Atrey, P. K., Kankanhalli, M. S., & Jain, R. Information assimilation framework for event detection in multimedia surveillance systems.
15. Oh, S., McCloskey, S., Kim, E., Vadat, A., Cannons, K. J., Hajimirsadeghi, H., et al. (2014). Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*.
16. Cricri1, F., Dabov, K., Curcio, I. D. D., Mate, S., & Gabbouj, M. (2014). Multimodal extraction of events and of information about the recording activity in user-generated videos. *Multimedia Tools and Applications*.



17. Koutras, P., Zlatinski, A., & Maragos, P. (2018). Exploring CNN-based architectures for multi-modal salient event detection in videos. In *Proceedings of the 13th IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) Workshop*.
18. Güder, M., & Çiçekli, N. K. (2018). Multimodal video event recognition based on association rules and decision fusion. *Multimedia Systems*.