Int J Multimed Info Retr
DOI 10.1007/s13735-012-0024-2

TRENDS AND SURVEYS

High-level event recognition in unconstrained videos

Yu-Gang Jiang · Subhabrata Bhattacharya · Shih-Fu Chang · Mubarak Shah

Received: 10 September 2012 / Accepted: 15 October 2012

Abstract The goal of high-level event recognition is to automatically detect complex high-level events in a given video sequence. This is a difficult task especially when videos are captured under unconstrained conditions by non-professionals. Such videos depicting complex events have limited quality control, and therefore, may include severe camera motion, poor lighting, heavy background clutter, and occlusion. However, due to the fast growing popularity of such videos, especially on the Web, solutions to this problem are in high demands and have attracted great interest from researchers. In this paper, we review current technologies for complex event recognition in unconstrained videos. While the existing solutions vary, we identify common key modules and provide detailed descriptions along with some insights for each of them, including extraction and representation of low-level features across different modalities, classification strategies, fusion techniques, etc. Publicly available benchmark datasets, performance metrics, and related research forums are also described. Finally, we discuss promising directions for future research.

Y.-G. Jiang (✉)
School of Computer Science, Fudan University, Shanghai, China
e-mail: ygj@fudan.edu.cn

S. Bhattacharya · M. Shah
Computer Vision Lab, University of Central Florida,
Orlando, FL, USA
e-mail: subh@cs.ucf.edu

M. Shah
e-mail: shah@eecs.ucf.edu

S.-F. Chang
Department of Electrical Engineering, Columbia University,
New York, NY, USA
e-mail: sfchang@ee.columbia.edu

Keywords Video events · Recognition · Unconstrained videos · Multimedia event detection · Multimodal features · Fusion

1 Introduction

High-level video event recognition is the process of automatically identifying video clips that contain events of interest. The high-level or complex events—by our definition—are long-term spatially and temporally dynamic object interactions that happen under certain scene settings. Two popular categories of complex events are instructional and social events. The former includes *procedural videos* (e.g., "making a cake", "changing a vehicle tire"), while the latter includes *social activities* (e.g., "birthday party", "parade", "flash mob"). Techniques for recognizing such high-level events are essential for many practical applications such as Web video search, consumer video management, and smart advertising.

The focus of this work is to address the issues related to high-level event recognition. Events, actions, interactions, activities, and behaviors have been used interchangeably in the literature [1,15], and there is no agreement on the precise definition of each term. In this paper, we attempt to provide a hierarchical model for complex event recognition in Fig. 1. Movement is the lowest level description: "an entity (e.g. hand) is moved with large displacement in right direction with slow speed". Movements can also be referred as attributes which have been recently used in human action recognition [73] following their successful use in face recognition in a single image. Next are activities or actions, which are sequences of movements (e.g. "hand moving to right followed by hand moving to left", which is a "waving" action). An action has a more meaningful interpretation and is often

Published online: 13 November 2012

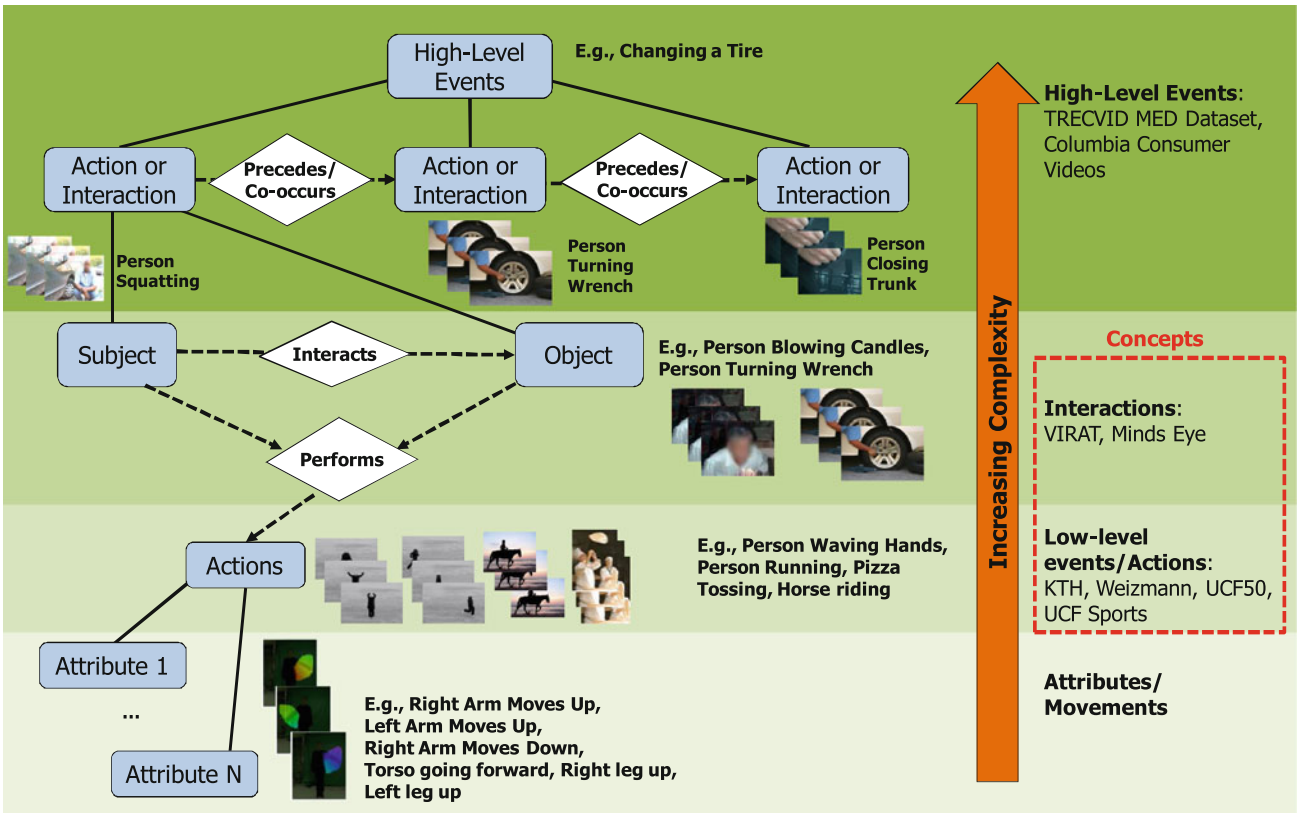

Fig. 1 A taxonomy of semantic categories in videos, with increased complexity from bottom to top. *Attributes* are basic components (e.g., movements) of *actions*, while actions are key elements of *interactions*. *High-level events* (focus of this paper) lie on top of the hierarchy, which contain (normally multiple) complex actions and interactions evolving over time

performed by entities (e.g., human, animal, and vehicle). An action can also be performed between two or more entities, which is commonly referred to as an *interaction* (e.g., person lifts an object, person kisses another person, car enters facility, etc.). Motion verbs can also be used to describe interactions. Recently the Mind's eye dataset is released under a DARPA program which contains many motion verbs such as "approach", "lift", etc [11]. In this hierarchy, *concepts* span across both actions and interactions. In general, *concept* is a loaded word, which has been used to represent objects, scenes, and events, such as those defined in large-scale concept ontology for multimedia (LSCOM) [95]. Finally, at the top level of the hierarchy, we have *complex* or *high-level* events that have larger temporal durations and consist of a sequence of interactions or stand-alone actions, e.g., an event "changing a vehicle tire" contains a sequence of interactions such as "person opening trunk" and "person using wrench", followed by actions such as "squatting" and so on. Similarly, another complex event such as "birthday party" may involve actions like "person clapping" and "person singing", followed by interactions like "person blowing candle" and "person cutting cake". Note that although we have attempted to encapsulate most semantic components of complex events in a single hierarchy, because of the polysemous nature of the words, adopting the same terminologies in the research community is an impossible objective to achieve.

Having said that, we set the context of event recognition as the *detection* of temporal and spatial locations of the complex event in the video sequence. In a simplified case when temporal segmentation of video into clips has been achieved, or where each video contains only one event and precise spatial localization is not important; it reduces to a video classification problem.

While many existing works have only employed the visual modality for event recognition, it is important to emphasize that video analysis is intrinsically multimodal, demanding multidisciplinary knowledge and tools from many fields, such as computer vision, audio and speech analysis, multimedia, and machine learning. To deal with large scale data that is common nowadays, scalable indexing methods and parallel computational platforms are also becoming an important part of modern video analysis systems.

There exist many challenges in developing automatic video event recognition systems. One well-known challenge is the long-standing semantic gap between computable low-level features (e.g., visual, audio, and textual features) and

semantic information that they encode (e.g., the presence of meaningful classes such as "a person clapping", "sound of a crowd cheering", etc.) [129]. Current approaches heavily rely on classifier-based methods employing directly computable features. In other words, these classifiers attempt to establish a correspondence between the computed features or a quantized layer on the features to the actual label of the event depicted in the video. In doing so, they lack a semantically meaningful, yet conceptually abstract, intermediate representation of the complex event, which can be used to explain what a particular event is, and how such representation can be used to recognize other complex events. This is why, with much progress made in the past decade in this context, the computational approaches involved in complex event recognition are reliable only under certain domain-specific constraints.

Moreover, with the popularity of handheld video recording devices, the issue becomes more serious since a huge amount of videos are currently being captured by non-professional users under unconstrained conditions with limited quality control (for example, in contrast to videos from broadcast news, documentary, or controlled surveillance). This amplifies the semantic gap challenge. However, it also opens a great opportunity since the proliferation of such user-generated videos has greatly contributed to the rapidly growing demands for new capabilities in recognition of high-level events in videos.

In this paper, we will first discuss the current popular methods for high-level video event recognition from multimedia data. We will review multimodal features, models, and evaluation strategies that have been widely studied by many groups in the recent literature. Compared to a few existing survey papers in this area (as summarized in the following subsection), this paper has a special focus on high-level events. We will provide in-depth descriptions of techniques that have been shown promising in recent benchmark evaluation activities such as the multimedia event detection (MED) task [99] of NIST TRECVID.[1] Additionally, we will discuss several important related issues such as the designs of evaluation benchmarks for high-level event recognition. Finally, we will identify promising directions for future research and developments. To stimulate further research, at the end of each important section we provide comments summarizing the issues with the discussed approaches and insights that may be useful for the development of future high-level event recognition systems.

[1] TREC video retrieval evaluation (TRECVID) [128] is an open forum for promoting and evaluating new research in video retrieval. It features a benchmark activity sponsored annually, since 2001, by the US National Institute of Standards and Technology (NIST). See http://trecvid.nist.gov for more details.

1.1 Related reviews

There have been several related papers that review the research of video content recognition. Most of them focused on human action/activity analysis, e.g., [1] by Aggarwal and Ryoo, [111] Poppe and [139] Turaga et al., where low-level features, representations, classification models, and datasets were comprehensively surveyed. While most human activity research was done on constrained videos with limited content (e.g., clean background and no camera motion), recent works have also shifted focus to the analysis of realistic videos such as user-uploaded videos on the Internet, or broadcast, and documentary videos.

In [130], Snoek and Worring surveyed approaches to multimodal video indexing, focusing on methods for detecting various semantic concepts consisting of mainly objects and scenes. They also discussed video retrieval techniques exploring concept-based indexing, where the main application data domains were broadcast news and documentary videos. Brezeale and Cook [17] surveyed text, video, and audio features for classifying videos into a predefined set of genres, e.g., "sports" or "comedy". Morsillo et al. [94] presented a brief review that focused on efficient and scalable methods for annotating Web videos at various levels including objects, scenes, actions, and high-level events. Lavee et al. [67] reviewed event modeling methods, mostly in the context of simple human activity analysis. A review more related to this paper is the one by Ballan et al. [8], which discussed features and models for detecting both simple actions and complex events in videos.

Different from the existing surveys mentioned above, this paper concentrates on the recognition of high-level complex events from multimedia data, such as those mentioned in Fig. 1. Many techniques for recognizing objects, scenes, and human activities will be discussed to the extent that is needed for understanding high-level event recognition. However, providing full coverage of those topics is beyond the scope of this work.

1.2 Outline

We first organize the research of high-level event recognition into several key dimensions (shown in Fig. 2), based on which the paper will be structured. While the design of a real system may depend on application requirements, key components like those identified (feature extraction and recognition model) are essential. These core technologies will be discussed in Sects. 2 and 3. In Sect. 4, we discuss advanced issues beyond simple video-level classification, such as temporal localization of events, textual recounting of detection results, and techniques for improving recognition speed and dealing with large-scale data. To help stimulate new research

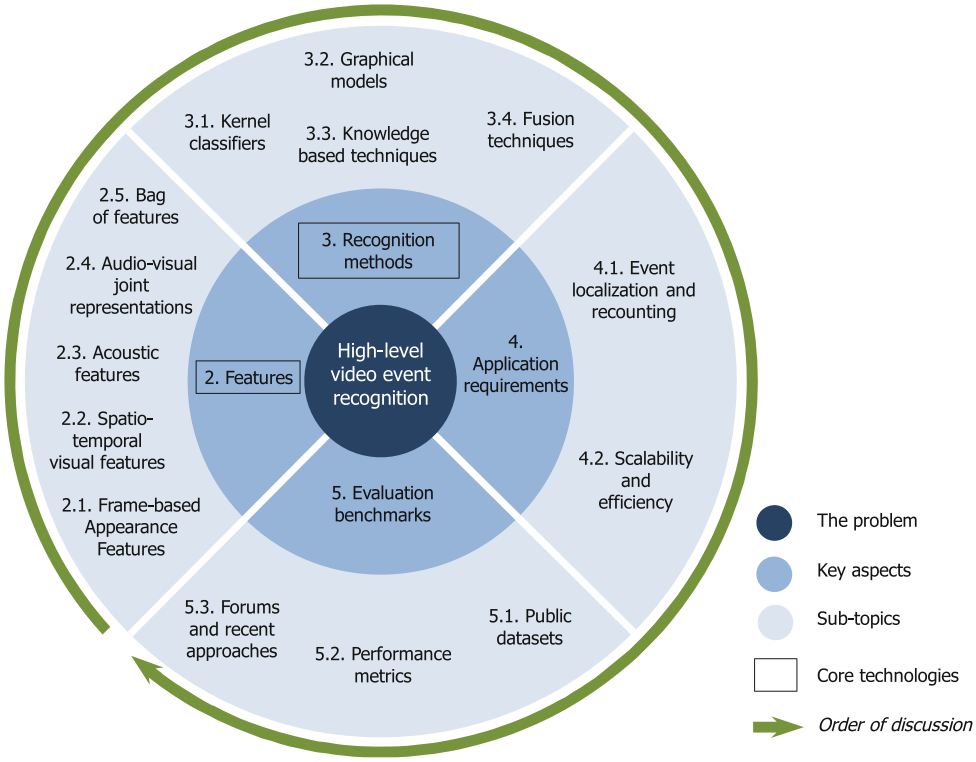

Fig. 2 Overview of various aspects of the video event recognition research. Presentation of the paper is structured based on this organization. Numbers correspond to the sections covering the topics

activities, we present reviews of popular benchmarks and explore insights of several top-performing systems in recent evaluation forums in Sect. 5. Finally, we discuss promising directions for future research in Sect. 6 and present conclusions in Sect. 7.

2 Feature representations

Features play a critical role in video event analysis. Good features are expected to be robust against variations so that videos of the same event class under different conditions can still be correctly recognized. There are two main sources of information that can be exploited. The visual channel, on one hand, depicts appearance information related to objects, scene settings, while on the other hand, captures motion information pertaining to the movement of the constituent objects and the motion of the camera. The second is the acoustic channel, which may contain music, environmental sounds and/or conversations. Both channels convey useful information, and many visual and acoustic features have been devised. We discuss static frame-based visual features in Sect. 2.1, spatio-temporal visual features in Sect. 2.2, acoustic features in Sect. 2.3, audio-visual joint representations in Sect. 2.4, and finally, the bag-of-features framework, which converts audio/visual features into fixed dimensional vectors in Sect. 2.5.

2.1 Frame-based appearance features

Appearance-based features are computed from a single frame. They do not consider the temporal dimension of video sequences but are widely used in video analysis since they are relatively easy to compute and have been shown to work well in practice. There has been a very rich knowledge base and extensive publicly available resources (public tools) devoted to static visual features. We divide existing works into local and global features, as will be discussed in the following.

2.1.1 Local features

A video frame can be represented efficiently using a set of discriminative local features extracted from it. The extraction of local features consists of two steps: detection and description. Detection refers to the process of locating stable patches that have some desirable properties which can be employed to create a "signature" of an image. In practice, uniform and dense sampling of image patches, with some obvious storage overhead is often used in comparison to the rather computationally expensive, less storage intensive patch detection [101].

Among popular local patch (a.k.a. interest point) detection algorithms, the most widely used one is Lowe's Difference-of-Gaussian (DoG) [77], which detects blob regions where

Fig. 3 Example results of local detectors: Harris–Laplace (*left*) and DoG (*right*). The images are obtained from [30]

the center differs from the surrounding area. Other popular detectors include Harris–Laplace [72], Hessian [88], maximally stable extremal regions (MSERs) [85], etc. Harris and Hessian focus on detection of corner points. MSER is also for blob detection, but relies on a different scheme. Unlike DoG which detects local maximums in multi-scale Gaussian filtered images, MSER finds regions whose segmentation is stable over a large range of thresholds. Figure 3 shows example results of two detectors. Interested readers are referred to [90] for a comprehensive review of several local patch detectors. Although it is observed that dense sampling eliminates the requirement of detectors, recent experiments shown in [34] confirm that methods using both strategies (sparse detection and dense sampling) offer the best performance in visual recognition tasks. In this direction, Tuytlaars proposed a hybrid selection strategy [140] where the author demonstrated how the advantages of both sampling schemes can be efficiently combined to improve recognition performance.

Once local patches are identified, the next stage is to describe them in a meaningful manner so that the resulted descriptors are (partially) invariant to rotation, scale, viewpoint, and illumination changes. Since the descriptors are computed from small patches as compared to a whole frame, they are also somewhat robust to partial occlusion and background clutter.

Many descriptors have been designed over the years. The best-known is scale-invariant feature transform (SIFT) [77], which partitions a patch into equal-sized grids, each described by a histogram of gradient orientations. A key idea of SIFT is that a patch is represented relative to its dominant orientation, which provides a nice property of rotation invariance. SIFT, coupled with several local detectors introduced above, has been among the most popular choices in recent video event recognition systems [10,58,96,98].

SIFT has been extended in various ways. PCA-SIFT was proposed by Ke et al. [60], who applied principal component analysis (PCA) to reduce the dimensions of SIFT. It stated that PCA-SIFT is not only compact but also more robust since PCA may help reduce noise in the original SIFT descriptors. However, such performance gains of PCA-SIFT were not found in the comparative study in [89]. An improved version of SIFT, called gradient location and orientation histogram (GLOH), was proposed in [89], to use a log-polar location grid instead of the original rectangular grid in SIFT. Work in [119] studied color descriptors that incorporated color information into the intensity-based SIFT for improved object and scene recognition. They reported a performance gain of 8 % on PASCAL VOC 2007 dataset.[2] Further, to improve the computational efficiency, Bay et al. [12] developed SURF as a fast alternative descriptor using 2D Haar wavelet responses.

Several other descriptors have also been popular in this context. Histogram of oriented gradients (HOG) was proposed by Dalal and Triggs [27] to capture edge distributions in images or video frames. Local binary pattern (LBP) [103] is another texture feature which uses binary numbers to label each pixel of a frame by comparing its value to that of its neighborhood pixels.

2.1.2 Global features

In earlier works global representations were employed, which encode a whole image based on the overall distribution of color, texture, or edge information. Popular ones include color histogram, color moments [166], and Gabor texture [83]. Oliva and Torralba [104] proposed a very low dimensional scene representation which implicitly encodes perceptual naturalness, openness, roughness, expansion, ruggedness using spectral and coarsely localized information. Since this represents the dominant spatial structure of a scene, it is referred to as the GIST descriptor. Most of these global features adopt grid-based representations which take spatial distribution of the scene into account (e.g., "sky" always appears above "road"). Features are computed within each grid separately and then concatenated as the final representation. This simple strategy has been shown to be effective for various image/video classification tasks.

Summary Single-frame based feature representations—such as SIFT, GIST, HOG, etc.—are the most straightforward to compute and have low-complexity. These features have been shown to be extremely discriminative for videos that do not depict rapid inter-frame changes. For videos with rapid content changes, one needs to carefully sample frames from which these features can be extracted if not all the frames are used. Since an optimal keyframe selection strategy is yet to be developed, researchers in practice sample frames uniformly. A low-sampling rate could lead to loss of vital information, while high sampling rates result in redundancies. Furthermore, these features do not include temporal information, and hence they are ineffective in representing motion, a very

[2] PASCAL visual object class (VOC) challenge is an annual benchmark competition on image-based object recognition, supported by EU-funded PASCAL2 Network of Excellence on Pattern Analysis, Statistical Modeling and Computational Learning.

important source of information in videos. This motivates us to move on to the next section that discusses spatio-temporal (motion) features.

2.2 Spatio-temporal visual features

Different from frame-based features, spatio-temporal features take the time dimension of videos into account, which is intuitively appealing since temporal motion information is critical for understanding high-level events.

2.2.1 Spatio-temporal local features

Many spatio-temporal video features have been proposed. Apart from several efforts in designing global spatio-temporal representations, a more popular direction is to extend the frame-based local features to work in 3D (x, y, t), namely spatio-temporal descriptors. In [65], Laptev extended the Harris corner patch detector [72] to locate spatial-temporal interest points (STIPs), which are space-time volumes in which pixel values have significant variations in both space and time. Figure 4 gives two example results of STIP detection. As will be discussed in later sections, STIP has been frequently used in recent video event recognition systems.

Several alternatives of STIP have been proposed. In Dollar et al. [29] proposed to use Gabor filters for 3D key-point detection. The detector, called Cuboid, finds local maxima of a response function that contains a 2D Gaussian smoothing kernel and 1D temporal Gabor filters. Rapantzikos et al. [112] used saliency to locate spatio-temporal points, where the saliency is computed by a global minimization process which leverages spatial proximity, scale, and feature similarity. To compute the feature similarity, they also utilized color, in addition to intensity and motion that are commonly adopted in other detectors. Moreover, Willems et al. [158] used the determinant of a Hessian matrix as the saliency measure, which can be efficiently computed using box-filter operations on integral videos. Wang et al. [151] conducted a comparative study on spatio-temporal local features and found that dense sampling works better than sparse detectors STIP, Cuboid, and Hessian, particularly on videos captured under realistic settings (in contrast to those taken in constrained environment with clean background). The dense sampling, however, requires a much larger number of features to achieve a good performance.

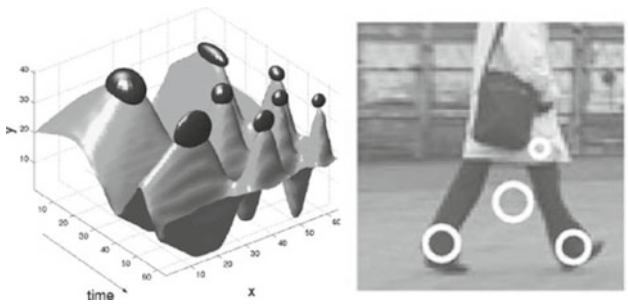

Fig. 4 Results of STIP detection using a synthetic sequence (*left*) and a realistic video (*right*; the detected points are shown on an image frame). The images are reprinted from [65] (©2005 Springer-Verlag)

Like the 2D local features, we also need descriptors to encode the 3D spatio-temporal points (volumes). Most existing 3D descriptors are motivated from those designed for the 2D features. Dollar et al. [29] tested simple flattening of intensity values in a cuboid around an interest point, as well as global and local histograms of gradients and optical flow. SIFT [77] was extended to 3D by Scovanner et al. [122], and SURF [12] was adapted to 3D by Knopp et al. [62]. Laptev et al. [66] used grid-based (by dividing a 3D volume into multiple grids) HOG and histogram of optical flow (HOF) to describe STIPs [65], and found the concatenation of HOG and HOF descriptors very effective. Biologically the combination of the two descriptors also makes good sense since HOG encodes appearance information while HOF captures motion clue. Klaser et al. [61] also extended HOG to 3D and proposed to utilize integral videos for fast descriptor computation. The self-similarities descriptor was adapted by Junejo et al. [123] for cross-view action recognition. Recently, Taylor et al.[135] proposed to use convolutional neural networks to implicitly learn spatio-temporal descriptors, and obtained similar human action recognition performance comparable to the STIP detector paired with HOG-HOF descriptors. Le et al. [69] combined independent subspace analysis (ISA) with ideas from convolutional neural networks to learn invariant spatio-temporal features. Better results from the ISA features over the standard STIP detector and HOG-HOF descriptors were achieved on several action recognition benchmarks [69].

2.2.2 Trajectory descriptors

Spatio-temporal information can also be captured by tracking the frame-based local features. These descriptors are theoretically superior to descriptors such as HOG-HOF, 3D SURF, Dollar Cuboids, etc. This is because they require the detection of a discriminative point or region over a sustained period of time, unlike the latter that computes various pixel-based statistics subjected to a predefined spatio-temporal neighborhood (typically $50 \times 50 \times 20$ pixels). However, computation of trajectory descriptors requires substantial computational overhead. The first of its kind was proposed by Wang et al. [149] where the authors used the well-known Kanade–Lucas–Tomasi (KLT) tracker [79] to extract DoG-SIFT key-point trajectories, and compute a feature by modeling the motion between every trajectory pair. Sun et al. [132] also applied KLT to track DoG-SIFT key-points.

Different from [149], they computed three levels of trajectory context, including point-level context which is an averaged SIFT descriptor, intra-trajectory context which models trajectory transitions over time, and inter-trajectory context which encodes proximities between trajectories. The velocity histories of key-point trajectories are modeled by Messing et al. [87], who observed that velocity information is useful for detecting daily living actions in high-resolution videos. Uemura et al. [141] combined feature tracking and frame segmentation to estimate dominant planes in the scene, which were used for motion compensation. In Yuan et al. [171] clustered key-point trajectories based on spatial proximities and motion patterns. Like [141], this method extracts relative features from clusters of trajectories on the background that describe the motion differently from those emanating from the foreground. As a result, the effect of camera motion can be alleviated using this approach. In addition, Raptis and Soatto [113] proposed tracklet, which differs from the long-term trajectories by capturing the local casual structure of action elements. A more recent work by Wu et al. [159] used Lagrangian particle trajectories and decomposed the trajectories into camera-induced and object-induced components, which makes their method robust to camera motion. Wang et al. [150] performed tracking on dense patches. They showed that dense trajectories significantly outperform KLT tracking of sparse key-points on several human action recognition benchmarks. In addition, a trajectory descriptor called motion boundary histogram (MBH) was also introduced in [150], which is based on the derivatives of optical flow. The derivatives are able to suppress constant motion, making MBH robust to camera movement. It has been shown to be very effective for action recognition in realistic videos [150]. Mostly recently, the work of [54] proposed to use local and global reference points to model the motion of dense trajectories, leading to a comprehensive representation that integrates trajectory appearance, location, and motion. The resulted representation is expected to be robust to camera motion, and also be able to capture the relationships of moving objects (or object-background relationships). Very competitive results were observed on several human action recognition benchmarks.

Summary Spatio-temporal visual features capture meaningful statistics from videos, especially those related to local changes or saliency in both the spatial and temporal dimensions. Most of the motion-based features are restricted to either optical flow or their derivatives. The role of semantically more meaningful motion features (e.g., kinematic features [2,4]) is yet to be tested in the context of this problem. Furthermore, most of these feature descriptors capture statistics based on either motion alone, or motion and appearance independently. Treating the motion and appearance modality jointly can further reveal important information which is lost in the process. Trajectories computed from local features have been shown to achieve performance gains at the cost of the computing overhead.

2.3 Acoustic features

Acoustic information is valuable for video analysis, particularly when the videos are captured under realistic and unconstrained environments. Mel-frequency cepstral coefficients (MFCC) is one of the most popular acoustic features for sound classification [7,33,163]. MFCC represents the short-term power spectrum of an audio signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. In Xu et al. [163] used MFCC together with another popular feature zeros crossing rate (ZCR) for audio classification. Predictions of audio categories such as "whistling" and "audience sound" are used for detecting high-level sports events like "foul", "goal", etc. Baillie and Jose [7] used a similar framework, but with MFCC features alone, for audio-based event recognition.

Eronen et al. [33] evaluated many audio features. Using a dataset of realistic audio contexts (e.g., "road", "supermarket" and "bathroom"), they found that the best performance was achieved by MFCC. In a different vein, Patterson et al. [107] proposed the auditory image model (AIM) to simulate the spectral analysis, neural encoding, and temporal integration performed by the human auditory system. In other words, AIM is a time-domain model of auditory processing intended to simulate the auditory images humans hear when presented with complex sounds like music, speech, etc. There are three main stages involved in the construction of an auditory image. First, an auditory filter bank is used to simulate the basilar membrane motion (BMM) produced by a sound in the cochlea (auditory portion of the inner ear). Next, a bank of hair cell simulators converts the BMM into a simulation of the neural activity pattern (NAP) produced at the level of the auditory nerve. Finally, a form of strobed temporal integration (STI) is applied to each channel of the NAP to stabilize any repeating pattern and convert it into a simulation of our auditory image of the sound. Thus, sequences of auditory images can be used to illustrate the dynamic response of the auditory image to everyday sounds. A recent work in this direction shows that features computed on auditory images perform better than the more conventional MFCC features for audio analysis [80].

Speech is another acoustic clue that can be extracted from video soundtracks. An early work by Chang et al. [22] reported that speech understanding is even more useful than image analysis for sports video event recognition. They used filter banks as features and simple template matching to detect a few pre-defined keywords such as "touchdown". Minami et al. [91] utilized music and speech detection to assist video analysis, where the detection is based on sound spectrograms. Automatic speech recognition (ASR) has also

been used for years in the annual TRECVID video retrieval evaluations [128]. A general conclusion is that ASR is useful for text-based video search over the speech transcripts but not for semantic visual concept classification. In [96], ASR was found to be helpful for a few events (e.g., narrative explanations in procedural videos), however not for general videos.

Summary Acoustic features have been found useful in high-level video event recognition. Although many new features have been proposed in the literature, currently the most popularly used one is still the MFCC. Developing new audio representations that are more suitable for video event recognition is an interesting direction.

2.4 Audio-visual joint representations

Audio and visual features are mostly treated independently for multimedia analysis. However, in practice, they are not independent except in some special cases where a video's audio channel is dubbed by an entirely different audio content, e.g., a motorbike stunt video dubbed with a music track. In the usual cases, statistical information such as co-occurrence, correlation, and covariance causality can be exploited across both audio and visual channels to perform efficient multimodal analysis.

In Beal et al. [13] proposed to use graphical models to combine audio and visual variables for object tracking. This method was designed for videos captured in a controlled environment and therefore may not be applicable to unconstrained videos. More recently, Jiang et al. [51] proposed a joint audio-visual feature, called audio-visual atom (AVA). An AVA is an image region trajectory associated with both regional visual features and audio features. The audio feature (MFCC of audio frames) and visual feature (color and texture of short term region tracks) are first quantized to discrete codewords separately. Jointly occurring audio-visual codeword pairs are then discovered using a multiple instance learning framework. Compared to simple late fusion of classifiers using separate modalities, better results were observed using a bag of AVA representation on an unconstrained video dataset [76]. This approach was further extended in [52], where a representation called audio-visual grouplet (AVG) was proposed. AVGs are sets of audio and visual codewords. The codewords are grouped together as an AVG if strong temporal correlations exist among them. The temporal correlations were determined using Granger's temporal causality [43]. AVGs were shown to be better than simple late fusion of audio-visual features.

The methods introduced in [51,52] require either frame segmentation or foreground/background separation, which is computationally expensive. Ye et al. [168] proposed a simple and efficient method called bi-modal audio-visual codewords. The bi-modal words were generated using normalized cut on a bipartite graph of visual and audio words, which capture the co-occurrence relations between audio and visual words within the same time window. Each bi-modal word is a group of visual and/or audio words that frequently co-occur together. Promising performance was reported in high-level event recognition tasks.

Summary Audio and visual features extracted using the methods discussed provide a promising representation to capture the multimodal characteristics of the video content. However, these features are still quite limited. For example, MFCC and region-level visual features may not be the right representation for discovering cross-modal correlations. In addition, the quality of the features may not be adequate due to noise, clutter, and motion. For example, camera motion can be a useful cue to discriminate between life events (usually depicting random camera jitter, zoom, pan, and tilt) and procedural events (usually static camera with occasional pan and/or tilt). None of the current feature extraction methods addresses this issue as the spatio-temporal visual feature extraction algorithms are not capable of distinguishing between the movement of the objects in the scene and the camera motion. Another disadvantage with feature-based techniques is that they are often ungainly in terms of both dimensionality and cardinality, which leads to storage issues as the number of videos is phenomenal. It is therefore desired to seek an additional intermediate representation for further analysis.

2.5 Bag of features

The local features (e.g., SIFT [77] and STIP [65]) discussed above vary in set size, i.e., the number of features extracted differs across videos (depending on complexity of contents, video duration, etc.). This poses difficulties for measuring video/frame similarities since most measurements require fixed-dimensional inputs. One solution is to directly match local features between two videos and determine video similarity based on the similarities of the matched feature pairs. The pairwise matching process is nevertheless computationally expensive, even with the help of indexing structures. This issue can be addressed using a framework called bag-of-features or bag-of-words (BoW) [127]. Motivated by the well-known bag-of-words representation of textual documents, BoW treats images or video frames as "documents" and uses a similar word occurrence histogram to represent them, where the "visual vocabulary" is generated by clustering a large set of local features and treating each cluster center as a "visual word".

BoW has been popular in image/video classification for years. The performance of BoW is sensitive to many implementation choices, which have been extensively studied in several works, mostly in the context of image classification with the frame-based local features like SIFT. Zhang

et al. [174] evaluated various local features and reported competitive object recognition performance by combining multiple local patch detectors and descriptors. Jiang et al. [55] conducted a series of analysis on several choices of BoW in video concept detection, including term weighting schemes (e.g., term frequency and inverse document frequency) and vocabulary size (i.e., the number of clusters). An important finding is that term weighting is very important, and a soft-weighting scheme was proposed to alleviate the effect of quantization error by softly assigning a descriptor to multiple visual words. The usefulness of such a soft weighting scheme is also confirmed in detecting a large number of concepts in TRECVID [21]. Similar idea of soft assignment was also presented by Philbin et al. [109]. van Gemert et al. [41] proposed an interesting approach called kernel codebooks to tackle the same issue of quantization loss. For vocabulary size, a general observation is that a few hundreds to several thousands of visual words might be sufficient for most visual classification tasks. In addition, Liu and Shah proposed to apply maximization of mutual information (MMI) for visual word generation [75]. Compared to typical methods like k-means, MMI is able to produce a higher level of word clusters, which are semantically more meaningful and also more discriminative for visual recognition. A feature selection method based on the page-rank idea was proposed in [74] to remove local patches that may hurt the performance of action recognition in unconstrained videos. Feature selection techniques were also adopted to choose discriminative visual words for video concept detection [56].

Spatial locations of the patches are ignored in standard BoW representation, which is not ideal, since the patch locations convey useful information. Lazebnik et al. [68] adopted a similar scheme like some of the global representations, by partitioning a frame into rectangular grids at various levels, and computing a BoW histogram for each grid. The histograms from grids at each level are then concatenated as a feature vector and pyramid match kernel is applied to measure the similarity between frames, each with multiple features from different spatial partitioning levels. This simple method has been proved effective in many applications and is now widely adopted. Researchers also found that direct concatenation of BoW histograms from grids of all levels plus support vector machines (SVMs) learning with standard kernels offers similar performance to the pyramid match kernel. However, it is worth noting that, different from BoW of the 2D frames, the spatial pyramid architecture has rarely been used in spatio-temporal feature-based BoW representations, again indicating the difficulty in handling the temporal dimension of videos.

The BoW framework can also be extended to represent a sound as a bag of audio words (a.k.a. bag-of-frames in the audio community), where the acoustic features are computed *locally* from short-term auditory frames (a time window of tens of milliseconds), resulting in a set of auditory descriptors from a sound. This representation has been studied in several works for audio classification, where the implementation choices slightly differ from that of the visual feature representations. Mandel et al. [82] used Gaussian mixture models (GMM) to describe a song as a bag of frames for classification. Aucouturier et al. [5] and Lee et al. [70] also adopted a similar representation. Aucouturier et al. conducted an interesting set of experiments and found that bag-of-frames performs well for urban soundscapes but not for polyphonic music. In place of GMM, Lu et al. [78] adopted spectral clustering to generate auditory keywords. Promising audio retrieval performance was attained using their proposed representation on sports, comedy, award ceremony, and movie videos. Cotton et al. [26] proposed to extract sparse transient features corresponding to soundtrack events, instead of the uniformly and densely sampled audio frames. They reported that, with fewer descriptors, transient features produce comparable performance to the dense MFCCs for audio-based video event recognition, and the fusion of both can lead to further improvements. The bag-of-audio-words representation has also been adopted in several video event recognition systems with promising performance (e.g., [10,58], among others).

Figure 5 shows a general framework of BoW representation, using different audio-visual features. A separate vocabulary is constructed for each feature type, by clustering the corresponding feature descriptors. Finally, a BoW histogram is generated for each feature type. Histograms can then be normalized to create multiple representations of the input video. In the simplest case, all the histogram representations can be concatenated to create a final representation before classification. This approach is usually termed as early fusion. An alternative approach is the late fusion where the histogram representations are independently fed into classifiers and decisions from the classifiers are combined. These will be discussed in detail later in Sect. 3.4.

Summary As it is evident that there is no single feature that is sufficient for high-level event recognition, current research strongly suggests the joint use of multiple features, such as static frame-based features, spatio-temporal features, and acoustic features. However, whether BoW is the best model to obtain meaningful representations of a video remains an important open issue. Although this technique performs surprisingly well [24,58,96,98], the major drawback of the systems conforming to this paradigm is their incapability to obtain deep semantic understanding of the videos, which is a prevalent issue in high-level event analysis. This is because, they provide a compact representation of a complex event depicted in a video based on the underlying features without having any understanding of the hierarchical

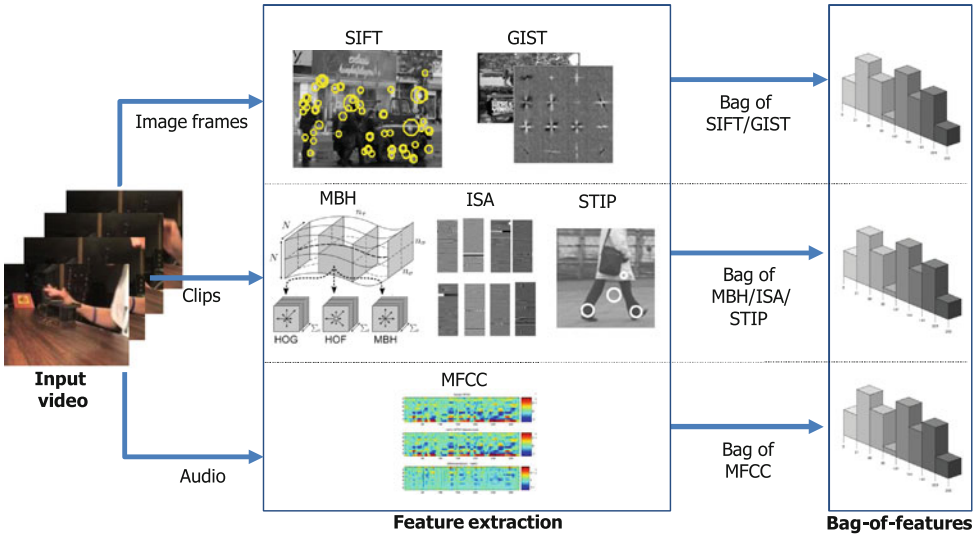

Fig. 5 Bag-of-features representations obtained from different feature modalities for high-level event detection

components, such as interactions or actions that constitute the complex event. Needless to say, the sense of spatio-temporal localization of these components is lost in this coarse representation. Besides, these methods also suffer from the usual disadvantages of quantization used in converting raw features to discrete codewords as pointed out in [16,109].

3 Recognition methods

Given the feature representations, event recognition can be achieved by various classifiers. This is a typical machine learning process, where a set of annotated videos are given for model training. The models are then applied to new videos for event recognition. We divide the discussion of recognition methods into four subsections. Section 3.1 introduces kernel classifiers, where we mainly discuss SVM, the most popular classifier in current event recognition systems. Section 3.2 discusses graphical methods, which are able to explicitly model temporal relationships between low-level events. Section 3.3 describes knowledge-based techniques, which can incorporate domain knowledge into event recognition. In Sect. 3.4, we discuss several fusion techniques to explore the power of combining multimodal features.

3.1 Kernel classifiers

Kernel-based classifiers have been popular in a wide range of applications for many years [45]. With kernels, linear classifiers that have been comprehensively studied can be applied in kernel space for nonlinear classification, which often leads to significantly improved performance. Among many choices of kernel-based classifiers (e.g., kernel Fisher discriminants), SVM is the most widely used algorithm due to its reliable performance across many different tasks, including high-level video event recognition. In the following sections, we discuss several issues related to applying SVM for video event recognition.

3.1.1 Direct classification

Event recognition is often formulated as a one-versus-all manner based on low-level representations, where a two-class SVM is trained to classify each event. For two-class SVM, the decision function for a feature vector $\mathbf{x}$ of a test video has the following form:

$$f(\mathbf{x}) = \sum_i \alpha_i y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) - b, \tag{1}$$

where $\mathcal{K}(\mathbf{x}_i, \mathbf{x})$ is the output of a kernel function for the feature of the ith training video $\mathbf{x}_i$ and the test sample $\mathbf{x}$; y_i is the event class label of $\mathbf{x}_i$; α_i is the learned weight of the training sample $\mathbf{x}_i$; and b is a learned threshold parameter.

Choosing an appropriate kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ is critical to the classification performance. For BoW representations of feature descriptors like SIFT or STIP, it has been reported that χ^2 Gaussian kernel is suitable [55,150,174], defined as

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\rho d_{\chi^2}(\mathbf{x},\mathbf{y})}, \tag{2}$$

where $d_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_j \frac{(x_j - y_j)^2}{x_j + y_j}$ is the χ^2 distance between samples $\mathbf{x}$ and $\mathbf{y}$. χ^2 Gaussian kernel was employed in all the recently developed top-performing high-level event recognition systems [10,48,58,96,98].

The performance of SVM classification is sensitive to a few parameters, among which the most critical one is ρ in the kernel function. The selection of a suitable parameter depends on data distribution, which varies from task to task. A common way is to use cross-validation, which evaluates a wide range of parameter values and picks the best one. However, this strategy is time-consuming. Recently, researchers

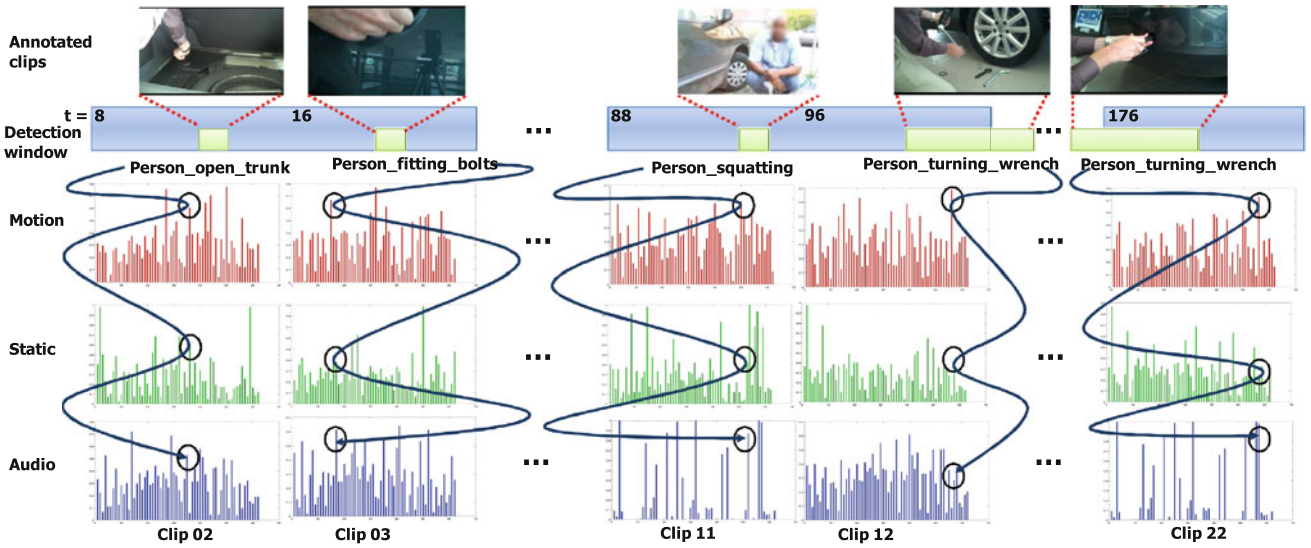

Fig. 6 Responses of lower level concept detectors in an arbitrary video depicting a complex event "changing a tire". This figure is best viewed in color. See texts for more explanations.

have empirically found that setting ρ as $1/\bar{d}$ often leads to near-optimal performance [174], where $\bar{d}$ is the mean of pairwise distances among all training samples. This simple strategy is currently widely adopted.

While accumulating all the feature descriptors from a video into a single feature vector seems a reasonable choice for event recognition, it neglects the temporal information within the video sequence. This issue can be addressed by using graphical models as will be described in Sect. 3.2. Another feasible solution is to use the earth mover's distance (EMD) [115] to measure video similarity. EMD computes the optimal flows between two sets of frames/clips, producing the optimal match between the two sets. Incorporating the EMD into a SVM classifier, the goal of temporal event matching can be achieved to a certain extent. This method was originally proposed by Xu et al. [162] for event recognition in broadcast news videos with promising results.

3.1.2 Hierarchical classification using concept-based recognition

Approaches under the *direct classification* category work satisfactorily to some extent. As discussed earlier, they are incapable of providing understanding of the semantic structure present in a complex event. Consider event "changing a vehicle tire", which typically consists of semantically lower level classes such as "person opening car trunk", "person using wrench", "person jacking car", etc. A bag of words representation collapses information into a long feature vector followed by direct classification is apparently not able to explain the aforementioned semantic structure.

This has motivated researchers to explore how an alternative representation could be efficiently utilized for semantic analysis of complex events. Events can be mostly characterized by several moving objects (person, vehicle, etc.), and generally occur under particular scene settings (kitchen, beach, mountain, etc.) with certain audio sounds (metallic clamor, wooden thud, cheering, etc.) and cues from overlaid or scene texts (street names, placards, etc.). Detection of these intermediate concepts has been proved to be useful for high-level event recognition.

Figure 6 gives an example of concept detection results in a video of event "changing a tire", where the top row shows sampled frames. The blue horizontal bar gives a sense of the temporal sampling window, on which pre-trained concept detectors are applied. The smaller green horizontal bars correspond to the actual granularity of the lower level concept classes (obtained from manual annotation). The bottom 3 rows show the detector responses from different feature modalities (each vertical bar indicates a concept). After combining the responses of concept detectors from different modalities, we observe that the concept "person opens trunk" is detected with maximum confidence in the shown window. This is very close to the ground truth. Similar trend is observed for other concepts like "person fitting bolts", "person squatting" and "person turning wrench", which are all very relevant to the event "changing a tire".

A few efforts have been devoted to the definition of a suitable set of concepts. One representative work is the LSCOM ontology [95], which defined 1,000+ concepts by carefully considering their utility for video retrieval, feasibility of automatic detection, and observability in actual datasets. In addition, several works directly adopted the WordNet ontology (e.g, ImageNet [28]). A simple and popular way to utilize these concepts in event recognition is to adopt a two-layer SVM classification structure [10,24,96,98], where

each model in the first layer detects a semantic concept, and a second-level model is used to recognize event classes using a representation based on the first-layer outputs as feature. All these works [10,96,98] have reported notable but small performance gains from this hierarchical classification approach, after fusing it with direct event classification using low-level features like SIFT and STIP.

Once such an intermediate representation is established, there are a variety of techniques that can be applied for complex event detection. The idea of two-layer classification can be extended to model a more sophisticated event model using co-occurrence and covariance of concepts. In order to further exploit temporal dependencies between the constituting concepts, graphical models can be employed. A detailed discussion on the use of graphical models will be provided in Sect. 3.2.

The concept-based hierarchical classification framework has several advantages over direct classification approaches using low-level features. As described earlier, this methodology decomposes a complex event into several semantically meaningful entities, where many of the lower level concepts may be easier to be detected since training samples of these concepts are relatively less diverse or noisy. Also, this framework can be extended to discover ad hoc events with few exemplars, if the lower level concepts can be reliably detected. Furthermore, hierarchical classification also paves a way for event recounting as detection of these concepts can provide detailed information. However, despite of all the advantages, there are also a few drawbacks. First, the concept detectors alone require intense training and they are not guaranteed to perform consistently across different datasets. Second, obtaining high quality training data is a challenging and time-consuming task.

Summary Research of hierarchical event classification has not been explored heavily. Considering that the current approaches can only improve direct classification by a small gain, we believe that there is still room for substantial improvement from this recognition paradigm. Currently, concepts such as human faces, pedestrians, and simple scenes/objects can be detected fairly reliably. We conjecture that as more concepts are reliably detected, and as more advanced hierarchical models are designed, much more significant performance gain will be achieved. In addition, current selections of mid-level concepts are still ad hoc. A systematic way in discovering relevant concepts and constructing suitable ontological structures among concepts is lacking.

3.2 Graphical models

There has been a plethora of literature over the last few decades which advocate the use of graphical models for the analysis of sequential data. Most approaches under this category combine insights from probability and graph theory to find structure in sequential data. These approaches can be broadly categorized into two sub-categories: directed graphical models and undirected graphical models. Popular methods of the former category include hidden Markov models (HMMs), Bayesian networks (BNs) and their variants. Markov random fields (MRFs), Conditional random fields (CRFs), etc. belong to the latter.

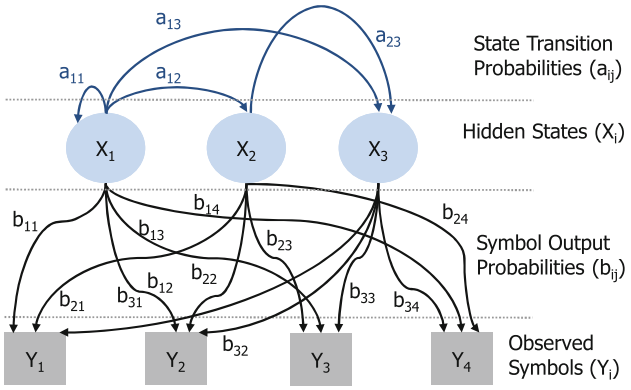

Fig. 7 An illustration of a typical discrete HMM. The model parameters can be obtained from model training

The simplest case of a directed graphical model is an HMM which adapts a single layer state-space formulation, wherein the outcome of an observed current state depends upon its immediately previous state. Observations can either be represented as discrete symbols (discrete HMM) or a continuous distribution (continuous HMM). A discrete HMM is explained in Fig. 7 where circular elements denote the hidden states, blue arrows denote the transitions between state pairs, gray rectangular elements are the observed symbols and the black arrows show the observation likelihood of a symbol given a state. Note that the directed arrows in the graph shown in Fig. 7 represent the transition between the hidden states and the observed states. In the context of complex event recognition, a directed graphical model is characterized by directed acyclic graphs which can be used to represent state-space relationships between constituent lower level events or sub-events.

The application of directed graphical models in activity or event recognition can be traced back to the work of Yamato et al. [164], where the authors proposed HMMs for recognizing tennis actions such as service, forehand volley, smash, etc. In their method, they extracted human figures by a standard background subtraction technique and binarized the resulting image. Mesh features on 8×8 binary patches were used to represent each image frame. These features were then transformed to a symbol sequence where each symbol encodes a keyframe in the input image sequence. For each action class, a separate discrete HMM was trained using the transformed symbol sequences.

Over the past 2 decades, several other works [71,97, 131,160] have used HMMs and their variants in human action recognition. Starner and Pentland [131] were among the early adopters of HMMs in their research on sign language recognition. Xie et al. [160] demonstrated how HMMs and hierarchical composition of multiple levels of HMMs could be efficiently used to classify play and non-play segments of soccer videos. Motivated by the success of HMMs, Li et al. [71] introduced an interesting methodology to model an action where hidden states in HMMs were replaced by visualizable salient poses (which forms an action) estimated using Gaussian mixture models. Since states in HMMs are not directly observable, mapping them to poses is an interesting idea. In the work by Natarajan and Nevatia [97], an action is modeled by a top-down approach, where the topmost level represents composite actions containing a single Markov chain, and the middle level represents primitive actions modeled using a variable transition HMM, followed by simple HMMs that form the bottommost layer representing human pose transitions. Recently, Inoue et al. [48] reported promising results in TRECVID MED task [99] using HMMs to characterize audio which is often observed to be a useful cue in multimedia analysis.

There are other types of directed graphical models that have been studied in event recognition. Another disadvantage with the HMM formulation is its incapability to model causality. This problem is alleviated by a different kind of directed graphical model called Bayesian networks (BN). BNs are capable of efficiently modeling causality using conditional independence between states. This methodology facilitates semantically and computationally efficient factorization of observation state space. In this vein, Intille and Bobick [49] introduced an agent-based probabilistic framework that exploits the temporal structure of complex activities typically depicted in American football plays. They used noisy trajectory data from soccer players collected from a static overhead camera to obtain temporal (e.g., before or after) and logical (e.g., pass or no pass) relationships, which are then used to model interactions between multiple agents. Finally, the BNs are applied to identify 10 types of strategic plays.

BNs cannot implicitly encapsulate temporal information between different nodes or states in the finite state machine model. Dynamic Bayesian networks (DBNs) can achieve this by exploiting the factorization principles available in Bayesian methods while preserving the temporal structure. Research on event recognition using DBNs is relatively new as compared to other approaches since it requires a certain amount of domain knowledge. Huang et al. [47] presented a framework for semantic analysis of soccer videos using DBNs, where they successfully recognized events such as corner kicks, goals, penalty kicks, etc.

BNs, HMMs and their variants fall under the philosophy of generative classification, which models the input, reducing variance of parameter estimation at the expense of possibly introducing model bias. Because of the generative nature of the model, a distribution is learned over the possible observations given the state. However, during inference or classification, it is the observation that is provided. Hence, it is more intuitive to condition on the observation, rather than the state.

This has motivated researchers to investigate alternative strategies for modeling complex events using undirected graphical models, some of which are naturally suited for discriminative modeling tasks. To this end, Vail et al. [144] made a strong contribution by introducing Conditional Random Fields for activity recognition. In their work, the authors show that CRFs can be discriminatively trained based on conditioning on the entire observation sequence rather than individually observed sample. A CRF can be perceived as a linear chain HMM without any directional edges between the hidden states and observations. In case of HMMs, the model parameters (transition, emission probabilities) are learned by maximizing the joint probability distribution, whereas, the parameters of a CRF (potentials) are learned by maximizing the conditional probability distribution. As a consequence, while learning the parameters of a CRF, modeling the distribution of the observations is not taken under consideration. The authors of [144] produced convincing evidence in favor of CRFs against HMMs in context of activity recognition. Inspired by the success of [144], Wang and Suter [153] introduced a variant of CRFs which can efficiently model the interactions between temporal order of human silhouette observations for complex event recognition. Wang and Mori [154] extended the idea of general CRFs to a max-margin hidden CRF for classification of human actions, where they model a human action as a global root template and a constellation of several "parts". More recently, in [25], Conolly proposed modeling and recognition of complex events using CRF, by taking observations obtained from multiple stereo systems under surveillance domain.

Although undirected graphical models (CRFs) are far less complex than their directed counterparts (DBNs), and avail all the benefits of discriminative classification techniques, they are disadvantageous in situations where the dependency between an event/action and its predecessors or successors (e.g., cause and effect) needs to be modeled. Although some variants of CRFs can overcome this problem by incorporating additional constraints and complex parameter learning techniques, they are computationally slow.

Summary Graphical models build a factorized representation of a set of independencies between components of complex events. Although the approaches discussed under this section are mathematically and computationally elegant,

their success in complex event recognition is still inconclusive. However, since these models provide an implicit level of abstraction in understanding complex events, research in this direction is expected to gather impetus as fundamental problems such as feature extraction and concept detection become more mature. With this said, we now move on to an alternative approach towards building this abstraction using knowledge-based approaches, in particular, techniques frequently employed in natural language processing.

3.3 Knowledge-based techniques

Knowledge-based techniques normally involve the construction of event models, and are usually used in special domains such as airport or retail surveillance, parking lot security and so on, where high-level semantic knowledge can be specified by the relevant domain experts. Let us consider the case of parking lot security. The usual events are mostly deterministic in nature, e.g., the "parking a vehicle" event would typically include the following sub-events: "vehicle enters garage through entry", "vehicle stopping near parking space", "person coming out of car", and "person exiting garage". The temporal and spatial constraints for each of these lower level concepts are known to a domain expert. For example, a vehicle can stop in the the driveway (spatial) for only a small amount of time (temporal). Thus, any violations to these specified constraints/knowledge would be considered as an outlier by the event model.

The knowledge of the context of an event is an extremely useful cue towards understanding the event itself. Researchers have extensively used domain knowledge to model events using different approaches. The work of Francois et al. [39] is noteworthy as the authors attempted to envision an open standard for understanding video events. One important part of the work is the modeling of events as composable, whereby complex events are constructed from simpler ones by operations such as sequencing, iteration, and alternation. In addition, the authors compiled an ontological framework for knowledge representation called video event representation language (VERL) based on foundations of formal language. In VERL, they described complex events by composing simpler primitive events, where sequencing is the most common composition operation. For example, an event involving a person getting out of a car and going into a building is described by the following sequence: opening car door, getting out of car, closing car door (optional), walking to building, opening building door, and entering building. The authors also provided an accompanying framework for video event annotations known as video event markup language (VEML).

Since a complex event can be formulated as a sequence of lower level or primitive concepts, production rules from formal language can be applied to generalize such complex events. Analogous to the language framework, where the concepts can be referred to as terminals while events can be named as non-terminals, the production rules can be augmented by probabilities of occurrence of terminals or non-terminals using stochastic context free grammar (SCFG) and their temporal relationships, e.g., a terminal precedes another using Allen's temporal algebra [3]. In this section, we summarize a few approaches that were proposed in the context of event recognition.

Ivanov and Bobbick [50] proposed a probabilistic syntactic approach to recognize complex events using the SCFG. The idea was integrated into a real-time system to demonstrate the approach in video surveillance application. Soon after, Moore and Essa [92] derived production rules using SCFG to represent multi-tasked activities as a sequence of object contexts, image features, and motion appearances from exemplars. In a similar note, Ryoo and Aggarwal [117] proposed the use of context free grammar, to represent an event as temporal processes consisting of poses, gestures, and sub-events. A specialized case of SCFG is attribute grammar, where the conditions on each production rule can be augmented with additional semantics from prior knowledge of the event domain. This has been used by Joo and Chellappa [59] who attempted to recognize atomic events in parking lot surveillance.

The influence of Case Grammar on contemporary linguistics has been significant. In linguistics, a case frame describes important aspects of semantic valency, of verbs, adjectives, and nouns [37]. For instance, the case frame of a verb "give" includes an Agent (A), an Object (O), and a Beneficiary (B), e.g., "NIST (A) gave video data (O) to the active participants (B)". This has inspired the development of frame-based representations in Artificial Intelligence. In [44], the authors extended the idea of case frame to represent complex event models. Case frame was extended to model the importance of causal and temporal relationships between low-level events. Multi-agent and multi-threaded events are represented using a hierarchical case frame representation of events in terms of low-level events and case-lists. Thus, a complex event is represented using a temporal tree structure, thereby formulating event detection as subtree pattern matching. The authors show two important applications of the proposed event representation for the automated annotation of standard meeting video sequences, and for event detection in extended videos of railroad crossings.

More recently, Si et al. [125] introduced AND-OR graphs to learn the event grammar automatically using a pre-specified set of unary (agent, e.g., person bending torso) and binary (agent-environment, e.g., person near the trash can) relations detected for each video frame. They demonstrated how the learned grammar can be used to rectify the noisy detection of lower level concepts in office surveillance.

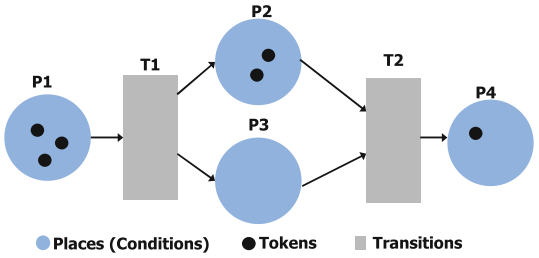

Fig. 8 An illustration of a typical place transition net

Efforts have also been made to represent and detect complex events based on first-order logic, generally known as Markov Logic Networks (MLNs). MLNs can be interpreted as graphs satisfying Markovian properties whose nodes are atomic formulas from first-order logic and the edges are the logical connectives used to construct the formulas. Tran and Davis [137] adopted MLNs to model complex interactions between people in parking lot surveillance scenarios by integrating common sense reasoning, e.g., a person can drive only one car at a time.

Knowledge representation can be achieved using networks or graphical structure. Ghanem et al. [42] used place transition networks or Petri Nets (PTNs) [20]. A Petri net provides an abstract model to represent the flow of information using a directed graphical model contrary to approaches that use undirected graphical models (e.g., HMMs), leading to a logical inferencing framework. PTNs can be explained with Fig. 8, where the hollow circles denote *places* containing solid circled *tokens*, rectangles depict transition and directed arrows are called *arcs* to show the direction of the flow. A change in the distribution of tokens inside a place node triggers a transition. This framework was used for event modeling by Cassel et al. [20] and later extended in [42] for parking lot surveillance. Here objects such as cars and humans were treated as tokens, single object or two object conditions such as moving/stationary and spatially near/far were considered the places, and primitive events such as start, stop, accelerate, and decelerate were the transitions between one place node to another. An example PTN model that exploits domain knowledge for counting the number of cars in a parking area, as given in [42], is to build a simple net linking primitive actions "Car C0 appears, Car C0 enters parking area, Car C0 stops, Car C0 leaves parking area" in a sequential order. During the inference process, the positions of tokens in the Petri net summarize the history of past events and predict what will happen in the future which facilitate incremental recognition over past events.

Summary Knowledge-based techniques, although easy to understand, make several assumptions which render them ineffective for event detection in unconstrained videos. As PTNs rely heavily on rule-based abstractions as opposed to probabilistic learning-based techniques, methods based on such formalism are too rigid to be applied to unconstrained cases where there are strong content diversities. Although MLNs incorporate rule-based abstraction in a probabilistic framework, there is no convincing evidence on whether the inferencing mechanism can handle complex scenarios where enumerating all possible rules is practically an infeasible task. For the same reason, representations discussed in [44,125] are not able to detect complex events in situations where (basically) no domain knowledge is available.

3.4 Fusion techniques

Fusing multiple features is generally helpful since different features abstract videos from different aspects, and thus may complement each other. To recognize complex events in unconstrained videos, acoustic features are potentially important since the original soundtracks of such videos are mostly preserved. This is in contrast to surveillance videos with no audio and broadcast/movie videos mostly with dubbed soundtracks, for which acoustic features are apparently less useful. We have briefly reviewed a few audio-visual representations in Sect. 2.4. In this section, we discuss techniques for fusing multiple visual and/or audio feature modalities.

The combination of multimodal features can be done in various ways. The most popular and straightforward strategies are early fusion and late fusion. As briefly described in Sect. 2.5, early fusion concatenates unimodal features into a long vector for event learning using kernel classifiers, while late fusion feeds each unimodal features to an independent classifier and fusion is achieved by linearly combining the outputs of multiple learners.

Sadlier and O'Connor [118] extracted several audio-visual features for event analysis in sports videos. The features were fused by early fusion. Sun et al. [133] extracted MFCC, SIFT, and HOG features for Web video categorization. Both early and late fusion were evaluated, and their results did not show a clear winner of the two fusion strategies. Since the early concatenation of features may amplify the "curse of dimensionality" problem, late fusion has been frequently adopted for multimodal event recognition in unconstrained videos [48,57,58,96]. SIFT, STIP, and MFCC features were lately fused by Jiang et al. [58] in their TRECVID 2010 MED system. Late fusion of a similar set of features was also used by Inoue et al. [48] for high-level event recognition.

In late fusion, the selection of suitable fusion weights is important. Equal weights (average fusion) were used in [57,58], while [48,96] adopted cross validation to select data adaptive weights optimized for different events and specific datasets. Using weighted late fusion, excellent event recognition results were achieved by [48,96] in the MED task of NIST TRECVID 2011 [99]. It is worth-noting that late fusion with adaptive weights generally outperforms average fusion when training and test data follow similar distribution. In case

there is a domain change between training and test videos, the weights learned from cross validation on the training set may not generalize well to the test set. Therefore, the choice of fusion strategies depends on specific application problem domains. In broad-domain Internet video analysis, adaptive weights are expected to be more effective according to the conclusions from recent developments.

In addition to early/late fusion, Tsekeridou and Pitas [138] took a different approach which combines audio-visual clues using interaction rules (e.g., person X talking in scene Y) for broadcast news video analysis. Duan et al. [31] used multiple kernel learning (MKL), which combines multimodal features at kernel level, for recognizing events in Internet videos. They also proposed a domain adaptive extension of MKL, to deal with data domain changes between training and test data, which often occur in Internet scale applications. MKL was also adopted in [132] to combine multiple spatio-temporal features computed on local patch trajectories.

Note that the fusion techniques discussed above are not restricted to any particular type of classifier. They can be combined across completely different classification strategies. For example, classifier confidences obtained from SVM using low-level feature representations can be fused with that obtained from other high-level classifiers (e.g., HMMs, DBNs, etc.) based on a completely different representation. The only issue that needs to be addressed while fusing outputs of classifier responses is that the classifier outputs need to be in the same confidence space. To deal with scale variations commonly seen in prediction scores from different classifiers, Ye et al. [169] proposed a rank-based fusion method that utilizes rank minimization and sparse error models to recover common rank orders of results produced by multiple classifiers.

Summary Current research in this direction is mostly limited to straightforward approaches of early or late fusion. However, to design a robust system for high-level event recognition, fusion techniques play an extremely important role. As research strives towards more efficient multimodal representation of videos, the study of better fusion techniques is expected to gather momentum.

4 Application requirements

In this section, we discuss several issues that have emerged due to application requirements, including event localization and recounting, and scalable techniques which are key to Internet scale processing.

4.1 Event localization and recounting

As discussed earlier, most works view event recognition as a classification process that assigns an input video a binary or a soft probability label according to the presence of each event. However, many practical applications demand more than video-level event classification. Two important problems that will significantly enhance fine-grained video analysis are spatial-temporal event localization and textual video content recounting. The former tries to identify the spatial-temporal boundaries of an event, while the latter aims at accurately describing video contents using concise natural languages. Technically, solutions to both problems may be only one step ahead of video classification, but they are still in their childhood, and may become mature with sufficient efforts paid in the next several years. We discuss them below.

4.1.1 Spatio-temporal localization

Direct video-level event recognition using kernel classifiers, where it is assumed that a video has already been temporally segmented into clips and the task of classifier is to assign each clip a label, has been extensively studied. However, locating exactly the spatial-temporal position where an event happens is relatively less investigated. One reason is that—for high-level complex events—it is difficult to define precise temporal boundaries, let alone the spatial positions. However, even knowing approximate locations could be beneficial for applications like precise video search.

Several efforts have been made to sports video event localization. Zhang and Chang [173] developed a system for event detection in baseball videos. Events can be identified based on score box detection and OCR (Optical Character Recognition). Since a baseball event follows a strict temporal order (e.g., it begins with a pitching view and ends with a non-active view), the temporal location can be easily detected. Xu et al. [161] proposed to incorporate web-casting text into sports video event detection and observed significant gain especially for the cases that cannot be handled by audio-visual features.

Besides sports events, several methods have been proposed for temporally localizing human actions in videos or spatially detecting objects (e.g., "person" and "car") in images. These techniques form a good foundation for future research of high-level event localization. Duchenne et al. [32] used movie script mining to automatically collect (noisy) training samples for action detection. To locate the temporal positions of human actions, they adopted a popularly used sliding window approach by applying SVM classification over temporal windows of variable lengths. Hu et al. [46] employed multiple instance learning to deal with spatial and temporal ambiguities in bounding-box-based human action annotations. This method was found useful when the videos are captured in complex scenes (e.g., supermarket). Similar to the idea of sliding window search, Satkin and Hebert [120] located the best segment in a video for action training by exhaustively checking all possible segments of the video.

Fig. 9 Video event recounting examples generated by the approach proposed by Tan et al. [134]. The one on the right is a failure case due to incorrect prediction of scene context (reprinted from [134], ©2011 ACM)

Oikonomopoulos et al. [102] proposed to learn action specific codebooks, where each codeword is an ensemble of local features, with spatial and temporal locations recorded. The learned codebook was used to predict the spatial-temporal locations of an action-of-interest in test videos, using a simple voting scheme with Kalman filter-based smoothing.

Spatially localizing objects in images has been extensively studied in the literature. One seminal work is the Viola-Jones real-time object detector [148], which is based on a boosted cascade learning framework, using features derived from integral images that can be computed more efficiently. Using a similar framework, Vedaldi et al. [145] integrated several features using multiple kernel learning, which led to one of the best-performing systems in the object detection task of 2009 PASCAL VOC Challenge. Among many other recent efforts on object detection, a representative work is by Felzenszwalb et al. [35], who used deformable part-based models with several important innovations like the proposal of a latent SVM formulation for model learning and strategies for selecting hard negative examples (those which are difficult to be differentiated). This approach is now popularly adopted.[3]

Spatio-temporal localization of concepts is helpful for localizing the occurrence of high-level events. Since concepts tend to co-occur spatio-temporally, once localization information of a key concept (e.g., human face) is available, the probabilities of detection of other co-occurring concepts increase, thereby enhancing the overall accuracy of event detection. However, this is a difficult task to achieve given the current stature of detectors. In addition to the difficulty of the task, the exhaustive search using typical sliding-window-based approaches add to the computational complexity of the detection algorithms, questioning their viability in practical recognition tasks.

4.1.2 Textual recounting

Multimedia event recounting (MER) refers to the task of automatic textual explication of an event depicted in a video. Recently, NIST introduced the MER task[4] whose goal is to produce a recounting that summarizes the key evidence of the detected event. Textual human-understandable descriptions of events in videos may be used for a variety of applications. Beyond more precise content search, event recounting can also enhance media access for people with low vision.

Kojima et al. [63] developed a system to generate natural language descriptions for videos of human activities. First, head and body movements were detected, and then case frames [37] were used to express human activities and generate natural language sentences. Recently, Tan et al. [134] proposed to use audio-visual concept detection (e.g., "baseball field" and "cheering sound") to analyze Internet video contents. The concept detection results are converted into textual descriptions using rule-based grammar. Some example results from their method are shown in Fig. 9.

Other relevant works on visual content recounting include [36,44,105,167], all of which focused primarily on images, however. Yao et al. [167] explored a large image database with region-level annotations for image parsing and results are then converted to textual descriptions using a natural language generation (NLG) technique called head-driven phrase structure grammar [110]. Ordonez et al. [105] used 1 million Flickr images to describe images. Their method is purely data-driven, i.e., a query image is described using descriptions of its most visually similar image in the database. In a similar spirit to [105], Feng and Lapata [36] also leveraged a large set of Internet pictures (mostly with captions), to automatically generate headline-like captions for news images. In order to produce short headline captions, they adopted a well-known probabilistic model of headline generation [9].

If events could be perfectly recognized, recounting might become an easier problem, since incorporating knowledge from text tags of similar videos would probably work. A more sophisticated approach is to employ hierarchical concept-based classification as discussed in Sect. 3.1.2, which can provide key evidences in support of the detected events to perform recounting. This can provide additional information

[3] Source codes from the authors of [35] are available at http://www.cs.brown.edu/~pff/latent/.

[4] http://www.nist.gov/itl/iad/mig/mer.cfm.

for recounting, in particular when used in conjugation with a model that captures the temporal dependency across concepts (e.g., HMMs). A reasonable recounting for a complex event such as "baking a cake" can be exemplified as follows with the key evidences italicized for legibility: This video is shot in an *indoor kitchen* environment. A *person points finger* to the *ingredients*. *Then* he *mixes* the *ingredients* in a *bowl* using a *blender* whose *noise is heard* in the background. *After that* he *puts* the *bowl* in a *convectional oven*. *Finally* he *takes the bowl* and *puts it* on a *table*.

4.2 Scalability and efficiency

Speed is always a challenge when dealing with Internet-scale data. Current video event recognition research normally deals with only a few hundreds to tens of thousands of videos. To scale up, extra care needs to be taken in choosing features that are efficient to be extracted, recognition models which are fast, as well as a system architecture suitable for parallel computing. In the following we briefly introduce several representative works on each of these aspects.

For features, the SURF descriptor [12] was proposed as a fast replacement of SIFT [88]. Knopp et al. [62] extended SURF to efficiently compute 3D spatio-temporal key-points. Further, several works have reported that dense sampling, which uniformly selects local 2D image patches or 3D video volumes, can be adopted in place of the expensive sparse keypoint detectors (e.g., DoG [88]) with a competitive recognition performance [101]. Uijlings et al. [142] observed that the dense SIFT and dense SURF descriptors can be computed more efficiently with careful implementations that avoid repetitive computations of pixel responses in overlapping regions of nearby image patches.

The quantization or word assignment process in the BoW representation [127] is computationally expensive using brute-force nearest neighbor search. Nister et al. [100] showed that quantization can be executed very efficiently if words in the vocabulary are organized in a tree structure. Moosmann et al. [93] adopted random forest, a collection of binary decision trees, for fast quantization. Shotton et al. [124] proposed semantic texton forests (STF) as an alternative image representation. STFs are ensembles of decision trees that work directly on image pixels, and therefore can be efficiently computed since they do not require expensive local key-point detection and description. Yu et al. [170] further extended STF for efficient 3D spatio-temporal human action representation.

As introduced earlier, currently the most popular recognition method is the SVM classifier. The classification process of SVM could be slow when nonlinear kernels such as histogram intersection and χ^2 are adopted. Maji et al. [81] proposed an interesting idea, with which the histogram intersection and χ^2 kernels can be computed with logarithmic complexity in the number of support vectors. Uijlings et al. [142] tested this method on video concept detection tasks and observed a satisfying performance in both precision and speed. Recently, Jiang [53] conducted an extensive evaluation of the efficiency of features and classifier kernels in video event recognition. The fast histogram intersection kernel was reported to be reliable and efficient. In addition, the simple and efficient linear kernel was shown to be effective on high-dimensional feature representations like the Fisher vectors [108].

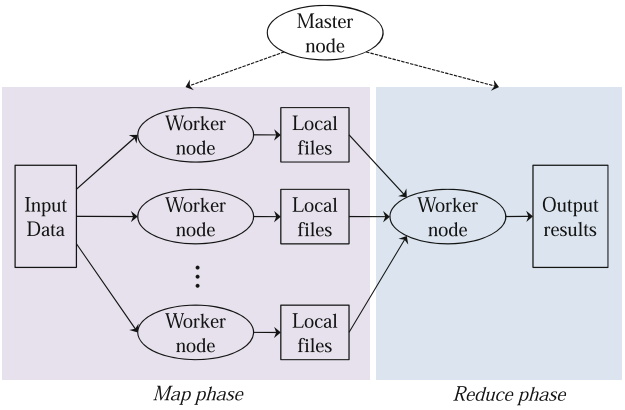

Fig. 10 Illustration of a typical MapReduce process

On the other hand, learning and inference algorithms for graphical models have been extensively investigated in the machine learning and pattern recognition fields. Frey and Jojic [40] evaluated several popular inference and learning algorithms of graph-based probability models in vision applications. de Campos and Ji [18] proposed an efficient algorithm which integrates several structural constraints for learning Bayesian Networks.

Parallel computing is very important for large-scale data processing. Video event recognition is not a task difficult to be split and run on multiple machines in parallel, as there could be many event categories, and each may be handled by one computer (node). In addition, testing videos can also be processed independently. MapReduce is probably the most popular framework for processing such a highly distributable problem. In MapReduce, a task is a basic computation unit such as classifying a video clip using a SVM model. Figure 10 depicts a general MapReduce process. The "Map" step employs a master node to partition the problem into several tasks and distribute them to worker nodes for computation. In the "Reduce" step, one or multiple worker nodes take the results and consolidate them to form the final output, which should be the same as running the entire problem on a single node. Several works have discussed the use of MapReduce in video processing. Yan et al. [165] adopted MapReduce for large-scale video concept recognition. They proposed a task scheduling algorithm specifically tailored

Table 1 Overview of TRECVID MED 2010–2011 [99] and CCV [57] datasets

Dataset	# Training/test videos	# Classes	# Positive videos per class	Average duration (s)	Format	File size (GB)
MED 2010	1,746/1,741	3	89	119	mp4	38
MED 2011	13,115/32,061	10	253	114	mp4	559
CCV	4,659/4,658	20	394	80	flv	30

The TRECVID videos are available upon participation of the benchmark evaluation, while the CCV dataset is publicly available. For all the three datasets, the positive videos are evenly distributed in the training and test sets

for the concept detection problem where the execution time varies across different tasks (e.g., classifying a video using SVM models with different number of support vectors). The algorithm estimates the computational time of each task a priori, which effectively compresses system idle time. White et al. [157] discussed MapReduce implementations of several popular algorithms in computer vision and multimedia problems (e.g., classifier training, clustering, and bag-of-features).

Another possible way to parallel event recognition algorithms is to use tightly coupled computational frameworks (computational modules make active communication with each other) such as message passing interface (MPI).[5] This approach, although more efficient, requires a total algorithmic redesign and a steep learning curve for multimedia and computer vision researchers. Therefore, a more practical solution is to use the MapReduce framework or other closely similar approaches such as the unstructured information management application (UIMA).[6] Since these approaches follow a loosely coupled computational paradigm where modules do not need to make active communication within themselves, they are expected to be favored by practitioners in the long run.

5 Evaluation benchmarks

Standard datasets for human action recognition research include those captured under constrained environments like KTH [121], Weizmann [14], IXMAS [155] and several more realistic ones such as UCF11 [74], UCF Sports [114], UCF50 action dataset [143], the Hollywood Movie dataset [66], and the more recently released Human Motion Database (HMDB) [64]. These benchmark datasets have played a very important role in advancing the state of the arts in human action analysis. In this section, we discuss evaluation benchmarks for high-level event recognition in unconstrained videos.

5.1 Public datasets

TRECVID MED datasets [99] Motivated by the need of analyzing complex events in Internet videos, the annual NIST TRECVID [128] activity defined a new task in 2010 called multimedia event detection (MED). Each year a new (or an extended) dataset is created for cross-site system comparison. Table 1 summarizes the 2010 and 2011 editions of TRECVID MED datasets. The MED data consist of user-generated content from Internet video hosting sites, collected and annotated by the Linguistic Data Consortium (LDC[7]). Figure 11 gives an example for each event class. In MED 2010, only three events were defined, all of which are long-term procedures. The number of classes increased to 15 in the much larger MED 2011 dataset. Out of the 15 classes, 5 are only annotated on the training set for system development (e.g., feature design and parameter tuning), and the remaining 10 are used in the official evaluation. Besides several procedure events, there are also a few social activity events included in 2011, e.g., "wedding ceremony" and "birthday party". The current editions of MED data only contain binary event annotations on video-level, and the MED task is focused only on video-level event classification.

Columbia consumer video (CCV) dataset [57] CCV[8] dataset was collected in 2011 to stimulate research on Internet consumer video analysis. Consumer videos are captured by ordinary consumers without professional post-editing. They contain very interesting and diverse content, and occupy a large portion in Internet video sharing activities (most of the MED videos are also consumer videos). A snapshot of the CCV dataset can be found in Table 1. 20 classes are defined, covering a wide range of topics including objects (e.g., "cat" and "dog"), scenes (e.g., "beach" and "playground"), sports events (e.g., "baseball" and "skiing"), and social activity events (e.g., "graduation" and "music performance"). Class annotations in CCV were also performed on video-level. The classes were defined according to the Kodak consumer video concept ontology [76]. The Kodak ontology contains over

[5] http://www.mcs.anl.gov/research/projects/mpi/.

[6] http://uima.apache.org/.

[7] http://www.ldc.upenn.edu/.

[8] Download site: http://www.ee.columbia.edu/dvmm/CCV/.

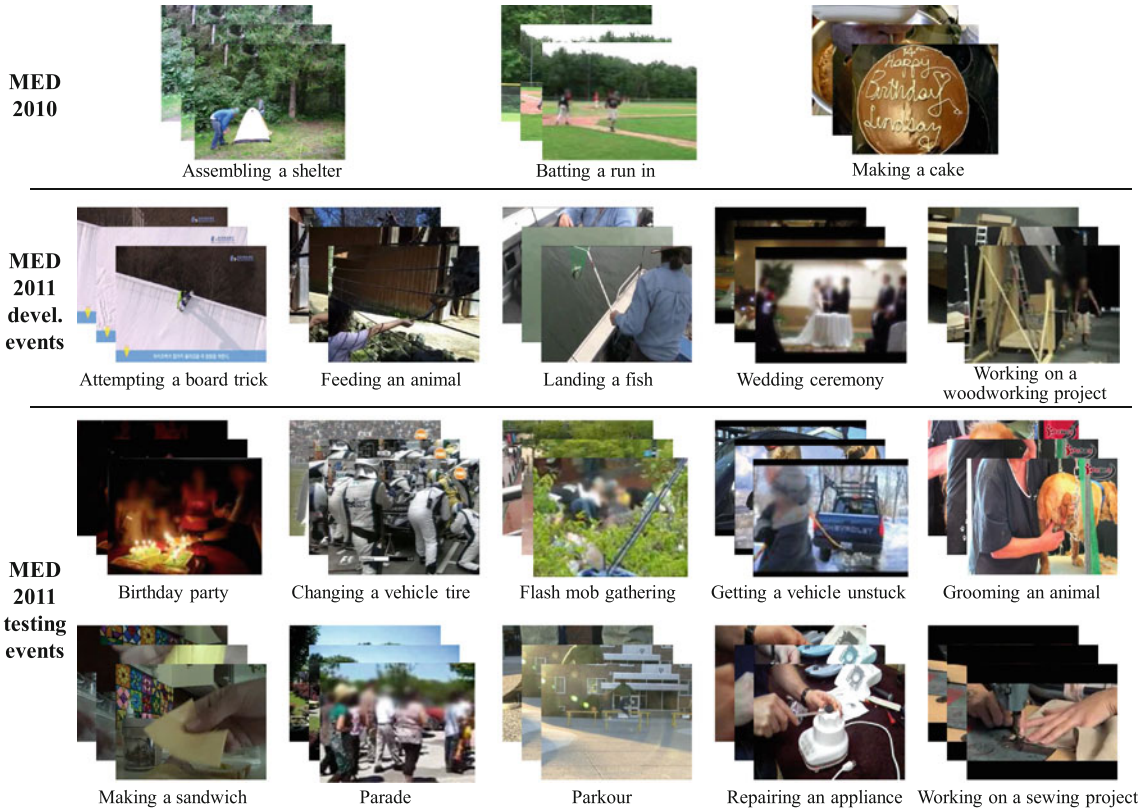

Fig. 11 Examples of TRECVID MED 2010 and 2011 events. In 2011, in addition to 10 events used for official evaluation, TRECVID also defined 5 events for system development (e.g, parameter tuning)

100 concept definitions based on rigorous user studies to evaluate the usefulness and observability (popularity) of each concept found in actual consumer videos.

Kodak consumer video dataset [76] Another dataset for unconstrained video analysis is the Kodak consumer video benchmark [76]. The Kodak consumer videos were collected by around 100 customers of Eastman Kodak Company. There are 1,358 video clips labeled with 25 concepts (a part of the Kodak concept ontology). Compared to MED and CCV datasets, one limitation of the Kodak dataset is that there is not enough intra-class variation. Many videos were captured under similar scenes (e.g., many "picnic" videos were taken at the same location), which make this dataset vulnerable to over-fitting issues.

There are also a few other datasets for unconstrained video analysis, e.g., LabelMe Video [172] and MCG-WEBV [19]. LabelMe Video is built upon the LableMe image annotation platform [116]. An online system is used to let Internet users to label not only event categories but also outlines and spatial-temporal locations of moving objects. The granularity of annotations is very suitable for finer-grained event recognition. However, since the labeling process is time-consuming and does not lead to any payment, the amount of collected annotations is dependent on highly motivated users. So far the annotations in LabelMe Video are quite limited in both scale and class diversity, and there is no video suitable for high-level event analysis. MCG-WEBV is a large set of YouTube videos organized by the Chinese Academy of Sciences. The current version of MCG-WEBV contains 234,414 videos, with annotations on several topic-level events like "a conflict at Gaza", which are too complicated and diverse to be recognized by content analysis alone. Existing works using this dataset are mostly for video topic tracking and documentation, by exploiting textual contexts (e.g., tags and descriptions) and metadata (e.g., video uploading time).

The availability of annotated data for training classifiers for event detection is a vital challenge. Recently, crowdsourcing efforts using the LabelMe toolkits [116,172] and the more general Amazon Mechanical Turk (AMT) platform[9] (used in the annotation of the CCV dataset) have been used extensively to annotate videos and images manually in a more efficient manner. It is expected to gain more popularity as researchers become aware of these tools.

5.2 Performance metrics

Event recognition results can be measured in various ways, depending on the application requirements. We first consider

[9] http://www.mturk.com/.

the most simple and popular case, where the determination of event presence is at the entire video level. This is essentially a classification problem: given an event of interest, a recognition system generates a confidence score for each input video.

Average precision (AP) and normalized detection cost (NDC) are the most widely used measurements of video event recognition performance. The input to both AP and NDC is a ranked video list according to the recognition confidence scores. We briefly introduce each of them in the following. In addition to AP and NDC, metrics based on detection-error tradeoff (DET) curves are being recently used to evaluate performance of event detection. The DET curves, as the name indicates, are generated from the probabilities of misclassification and false alarms produced by a given classifier.

Average precision AP is a single-valued measurement approximating the area under a precision-recall curve, which reflects the quality of the ranking of test videos (according to classification probability scores). Denote R as the number of true relevant videos in a target dataset. At any index j, let R_j be the number of relevant videos in the top j list. AP is defined as

$$\mathrm{AP} = \frac{1}{R}\sum_j \frac{R_j}{j} \times I_j,$$

where $I_j = 1$ if the jth video is relevant and 0 otherwise. AP favors highly ranked relevant videos. It returns a full score (AP = 1) when all the relevant videos are ranked on top of the irrelevant ones.

Normalized detection cost Figure 12 illustrates the basic concepts involved in computing the NDC, which is the official performance metric of the TRECVID MED task [38]. Different from AP that evaluates the quality of a ranked list, NDC requires a recognition threshold. Videos with confidence scores above the threshold are considered relevant (i.e., the relevant set in the figure). Specifically, given a recognition threshold, we first define P_{MD} (miss detection rate) and P_{FA} (false alarm rate):

$$P_{\mathrm{MD}} = \frac{\#misses}{\#targets},$$

$$P_{\mathrm{FA}} = \frac{\#false\ alarms}{\#total\ videos - \#targets},$$

where $\#targets$ is the total number of videos containing the target event in a dataset. With P_{MD} and P_{FA}, NDC can be computed as:

$$\mathrm{NDC} = \frac{C_{\mathrm{MD}} \times P_{\mathrm{MD}} \times P_T + C_{\mathrm{FA}} \times P_{\mathrm{FA}} \times (1 - P_{\mathrm{T}})}{\min(C_{\mathrm{MD}} \times P_{\mathrm{T}}, C_{\mathrm{FA}} \times (1 - P_{\mathrm{T}}))},$$

where P_{T} is the prior probability of the event (i.e., $\frac{\#targets}{\#total\ videos}$); C_{MD} and C_{FA} are positive cost parameters to weigh the importance of P_{MD} and P_{FA}, respectively.

As can be seen, NDC uses two cost parameters to weigh the importance of miss detection rate and false alarm rate. As a result, NDC provides a more flexible way than AP to evaluate recognition performance. Different from AP, lower NDC value indicates better performance. Based on NDC, NIST uses two variants to measure the performance of MED systems, namely ActualNDC and MinimalNDC. ActualNDC is based on the threshold provided by the participants based on their algorithms, while MinimalNDC is computed by the optimal threshold, i.e., the threshold that leads to the minimum NDC value on a ranked list. MinimalNDC is adopted as an additional measurement to ActualNDC since the latter is sensitive to the automatically predicted threshold.

Partial area under DET curve The DET curve, introduced by Martin et al. [84], is often used for evaluating detection performance where the number of negative samples is significantly larger than that of the positive ones. The curve is generated by plotting false alarm rate versus miss detection rate after scaling the axes non-linearly by their standard normal deviates. In order to quantitatively evaluate the performance of a classifier, the area under the DET curve can be used as a single-value metric which is inversely proportional to the classifier performance. However, the whole area under the curve may not be meaningful, which is why a portion of the curve under a predefined operating region is considered. Figure 13 illustrates the idea of using the partial area under DET curve as a metric under 60 % miss detection at 5 % false alarm operating region.

Spatio-temporal localization Unlike the video-level classification, spatio-temporal localization demands an evaluation measure that works in a finer resolution. Prior works on spatial [145] and temporal [32] localization are also evaluated by average precision (AP). Take temporal event localization as an example [32], systems return a list of video clips with variable durations (instead of a list of videos), ranked by the likelihood of *fully* containing the target event with no redundant frames. A clip is treated as a correct hit if it overlaps with a ground-truth event over a certain percentage (normally

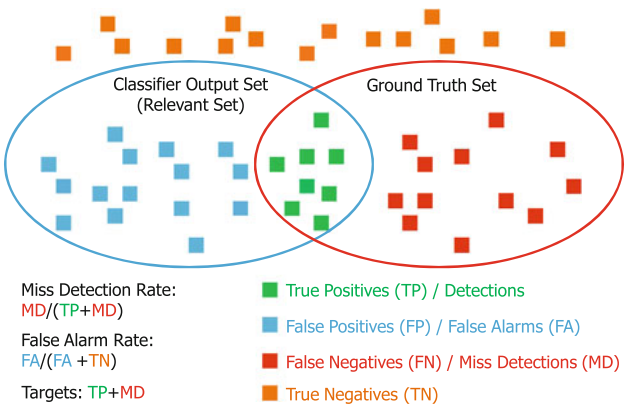

Fig. 12 An illustration of the terminologies used to compute NDC

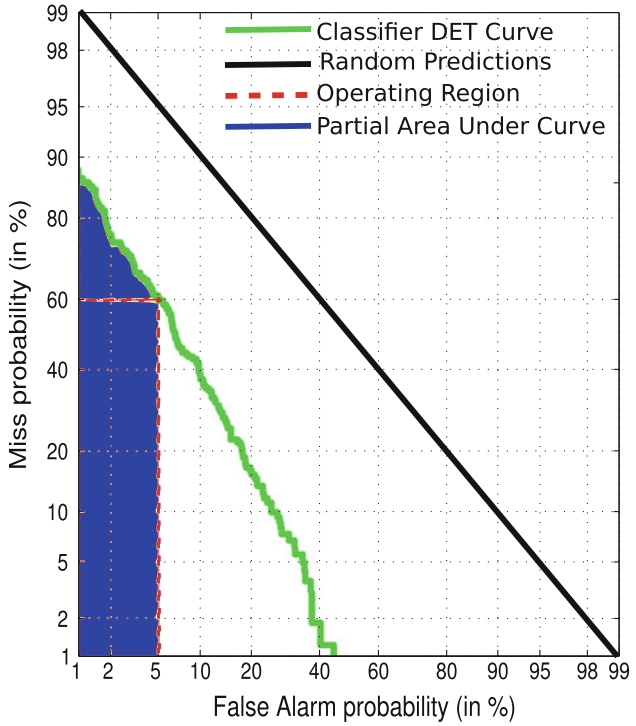

Fig. 13 An illustration of metric selection in a detection error tradeoff curve

50 %). With this judgment, AP can be easily applied over the ranked list of video clips. Similarly, this can be extended to spatial localization and spatial-temporal joint localization.

Multimedia event recounting The more challenging problem is event recounting, which is very difficult to be quantitatively evaluated. Most works on textual recounting used subjective user studies [134,105]. A number of criteria such as completeness and redundancy are defined, based on which users are asked to score each criterion. Some quantitative measures were also used as an additional measure to the subjective user evaluation. BLEU score [106], which is very popular for evaluating machine translation quality, was adopted by Ordonez et al. [105] to measure the quality of image captioning. Such a quantitative criterion, however, cannot really measure the consistency between the semantic meanings of the machine-generated sentences and the video content.

5.3 Forums and recent approaches

A few forums have been set up to stimulate video content recognition, e.g., NIST TRECVID [128] and MediaEval [86]. In this section, we focus our discussions on the annual NIST TRECVID evaluation [128], since it is to our knowledge the only forum that fully focuses on video analysis and has made consistently high impacts over the years. NIST defines several tasks every year, focusing on various issues on video retrieval. Among them, MED is the task evaluating systems for high-level event recognition in unconstrained videos. Initiated in 2010, MED already found its way to advance the state of the arts.

The number of participated teams in MED task has increased quickly from 7 (2010) to 19 (2011). We summarize all the submitted results in Fig. 14. Each team may submit multiple results to test the effectiveness of various combinations of system modules. The number of submitted results per team was limited to 4 in 2011. Such a limitation did not exist in 2010, and thus the number of submitted results did not increase at the same pace with the number of teams. In terms of mean MinimalNDC over the evaluated event classes, we see significant improvements from MED 2011 compared to the previous year. However, it is important to notice that results are not directly comparable across multiple years due to the changes in video data and event classes.

We briefly discuss the techniques of the teams who produced top-performing results. In 2010, the Columbia-UCF joint team [58] achieved the best performance using a framework combining multiple modalities, concept-level context (based on 21 scene/action/audio concept classifiers), and temporal matching techniques. Three audio-visual features (SIFT [77], STIP [65], and MFCC) were extracted and con-

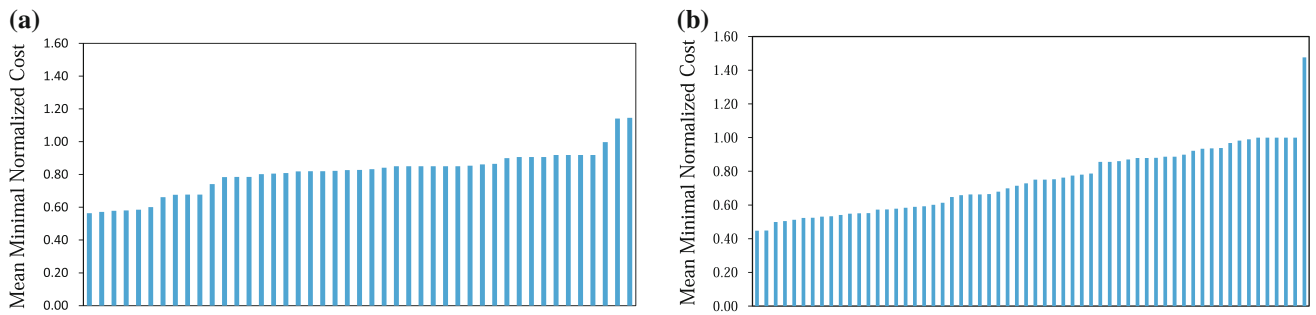

Fig. 14 Performance of TRECVID MED 2010 and 2011 submissions, measured using mean MinimalNDC over all the evaluated events. (a) MED 2010, 45 submissions from 7 teams. (b) MED 2011, 60 submissions from 19 teams. There are 3 test events in MED 2010 and 10 test events MED 2011

verted into the bag-of-words representations. SVM classifier was adopted to train models separately using each feature, and results were combined by average late fusion. Specifically, for SVM they used both standard χ^2 kernel and the EMD (earth mover's distance) kernel [162], where the latter was applied to alleviate the effect of event temporal misalignment. One important observation in [58] is that the three audio-visual features are highly complementary. While temporal matching with EMD kernel led to noticeable gain to some events, the concept-level context did not show clear improvements.

In MED 2011, the best results were achieved by a large collaborative team, named VISER [96]. Many features were adopted in the VISER system, such as SIFT [77], Color-SIFT [119], SURF [12], HOG [27], MFCC, Audio Transients [26], etc. Similar to [58], bag-of-words representation was used to convert each of the feature sets into a fixed-dimensional vector. A joint audio-visual bi-modal representation [168] was also explored, which encodes local pattern across the two modalities. Different fusion strategies were used—a fast kernel-based method for early fusion, a Bayesian model combination for optimizing performance at a specific operation point, and weighted average fusion for optimal performance over the entire performance curve. In addition, they also utilized other ingredients like object and scene level concept classifiers (e.g., the models provided in [136]), automatic speech recognition (ASR), and OCR. Their results showed that the audio-visual features, including the bi-modal representation, are very effective. The concept classifiers and the fusion strategies also offered some improvements, but the ASR and OCR features were less helpful perhaps due to their low occurrence frequencies in this specific dataset.

6 Future directions

Although significant efforts have been devoted to high-level event recognition during the past few years, the current recognition accuracy for many events is still far from satisfactory. In this section, we discuss several promising research directions that may improve event recognition performance significantly.

Better low-level features There have been numerous works focusing on the design of low-level features. Representative ones like SIFT [77] and STIP [65] have already greatly improved recognition accuracy, compared with the traditional global features like color and texture. However, it is clear from the results of the latest systems that these state-of-the-art low-level features are still insufficient for representing complex video events. Such handcrafted features, particularly the gradient-based ones (e.g., SIFT, HOG [27], and variants), are already reaching their limit in image and video processing. Thus, the community needs good alternatives that can better capture key characteristics of video events.

In place of the *handcrafted* static or spatial-temporal local features, a few recent works which exploited deep learning methods to automatically learn hierarchical representations [69,135,146] open up a new direction that deserves further studies. These automatically learned features already show similar or even better performance than the handcrafted ones on popular benchmarks. In addition to the visual features, another factor that should never be neglected is the audio track of videos, which is very useful as discussed earlier in this paper. Since audio and vision were mostly separately investigated in two different communities, limited research (except [51,168]) has been done on how audio-visual cues can be jointly used to represent video events (cf. Sect. 2.4). The importance of this problem needs to be highlighted to attract more research attention. We believe that good joint audio-visual representations may lead to a big leap in video event recognition accuracy.

Beyond BoW + SVM Most of the currently well performing event recognition systems rely on a simple pipeline that uses BoW representations of various visual descriptors and SVM classification. Although this approach, with years of study in optimizing the design details, has to-date the highest accuracy, the room for further improvement is very limited. Thus a natural question that arises is: Are there any more promising alternative solutions? While the exact solution may be unclear, the answer to the question is quite positive. There has been a recent surge in neural networks research on improving the accuracy of bag-of-words based representations [23,147]. These approaches show promising improvements in document classification over regular bag-of-words based approaches and hence are expected to improve event detection using conventional bag-of-words representation. Another interesting direction is to explore solutions that use prior knowledge, an intuitively very helpful resource that has been almost fully ignored in the current BoW + SVM pipeline. As it is true for humans that external knowledge is always important for perception, we believe it is also critical for the design of a robust automatic event recognition system. Although current knowledge-based models have not yet shown promising results, this direction deserves more investigation.

Event context and attributes Complex events can be generally decomposed into a set or sequence of concepts (actions, scenes, objects, audio sounds, etc.), which are relatively easier to be recognized since they have much smaller semantic granularity and thus are visually or acoustically more distinct. Once we have a large number of contextual concept detectors, the detection results can be applied to infer the existence of an event. As discussed earlier in Sect. 3.1.2, there are several works exploring this direction with, nevertheless, very straightforward modeling methods. In computer

vision, a similar line of research, namely attribute-based methods, also emerged recently for various visual recognition tasks. A few questions still need to be addressed: Whether one should manually specify concepts or attributes (supervised learning), or automatically discover them from an existing vocabulary (unsupervised learning)? How many and what concepts should be adopted? Is there a universal vocabulary of concepts that can be used for applications in any domain? How to reliably detect these concepts, and how to model events based on the concepts? Each of these problems requires serious and in-depth investigations. This may look like a difficult direction. However, once these problems are tackled, recognizing complex events would eventually be much more solvable.

Ad hoc event detection Ad hoc event detection refers to the cases where very few examples are available and the system does not have any prior knowledge about a target event. Techniques for ad hoc event detection is needed in retrieval-like scenarios, where users supply one or a few examples of an event-of-interest to retrieve relevant videos in a limited amount of time. Such problems are often termed as one-shot or few-shot learning. Apparently the knowledge-based solutions are incapable of performing this task since the event is not known a priori. The performance of supervised learning classifiers is questionable due to the small number of training samples. To this end, one can leverage knowledge from text to derive semantic similarity between annotated and undiscovered concepts, which can lead to the discovery of new concepts, for which there is no training data available [6]. The idea of semantic similarity can be extended to different levels of the event hierarchy to detect concepts with previously unseen exemplars, or complex events with no training exemplars. Following the discussions on event context, once the videos are offline indexed with a large number of concepts, online retrieval or detection of unknown events becomes possible since videos of the same event are very likely to contain similar concept occurrence distributions. In other words, event detection can be achieved by measuring the similarity of the concept occurrence vectors between query examples and database videos. This converts the ad hoc event detection task into a nearest neighbor search problem, to which highly efficient hashing techniques [152,156] or indexing methods [126] may be applied to achieve real-time retrieval in large databases.

Better event recounting Very limited works have been done on event recounting, although this capability is needed by many applications as discussed earlier. Precise video event recounting is a challenging problem that demands not only highly accurate content recognition but also good NLP models to make the final descriptions as natural as possible. Recognizing a large number of concepts (organized in a hierarchy) is certainly a good direction to pursue, where an interesting sub-problem is "how to *rectify* false detections based on contextual relationships (e.g., co-occurrence, causality, etc.) that exist across concepts?" To generate good descriptions, purely analyzing video content may not be sufficient for automatic techniques, which still have a long way to go to really reach humans' capability. To narrow this gap, the rich information on the Web may be a good complement since surrounding texts of visually similar videos may be exploited to recount a target video, even when the semantic content of the video cannot be perfectly recognized.

Better benchmark datasets The TRECVID MED task has set up a good benchmark for video event recognition. However, currently the number of events is still limited to 10–20, which is much fewer than the actual number of events that may appear in videos. On the one hand, this prevents the exploration of techniques that utilizes the co-occurrence or casuality between multiple events in a video. On the other hand, conclusions drawn from a small set of events may not generalize well. Therefore, there is a need to construct benchmark datasets covering a larger number of events. In addition, for event recounting there is still no well-defined datasets. To advance technical development in this direction, good datasets are desired.

7 Conclusions

We have presented a comprehensive survey of techniques for high-level video recognition in unconstrained videos. We have reviewed several important topics, including static frame-based features, spatio-temporal features, acoustic features, audio-visual joint representations, bag-of-features, kernel classifiers, graphic models, knowledge-based techniques, and fusion techniques. We also discussed several issues that emerged because of particular application requirements, such as event localization and recounting, as well as scalability and efficiency. Moreover, we described popular benchmarks and evaluation criteria, and summarized key components of systems that achieved top performance in recent TRECVID evaluations. With a few promising directions for future research given at the end, we believe that this paper can provide valuable insights for current researchers in the field and useful guidance for new researchers who are just beginning to explore this topic.

Acknowledgments This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract numbers D11PC20066, D11PC20070, and D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. Yu-Gang Jiang was partially supported by grants from the National Natural Science Foundation of China (#61201387 and #61228205).

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43(3):1–16
2. Ali S, Shah M (2010) Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans Pattern Anal Mach Intell 32(2):288–303
3. Allen JF (1983) Maintaining knowledge about temporal intervals. Commun ACM 26(11):832–843
4. Atkeson CG, Hollerbach JM (1985) Kinematic features of unrestrained vertical arm movements. J Neurosci 5(9):2318–2330
5. Aucouturier JJ, Defreville B, Pachet F (2007) The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. J Acoust Soc Am 122(2):881–891
6. Aytar Y, Shah M, Luo J (2008) Utilizing semantic word similarity measures for video retrieval. In: Proceedings of IEEE conference on computer vision and pattern recognition, Providence, USA
7. Baillie M, Jose JM (2003) Audio-based event detection for sports video. In: Proceedings of international conference on image and video retrieval, Urbana-Champaign, IL
8. Ballan L, Bertini M, Bimbo AD, Seidenari L, Serra G (2011) Event detection and recognition for semantic annotation of video. Multimedia Tools Appl 51(1):279–302
9. Banko M, Mittal VO, Witbrock, MJ (2000) Headline generation based on statistical translation. In: Proceedings of the annual meeting of the association for computational linguistics, Hong Kong
10. Bao L, Yu SI, Lan ZZ, Overwijk A, Jin Q, Langner B, Garbus M, Burger S, Metze F, Hauptmann A (2011) Informedia @ TRECVID 2011. In: Proceedings of NIST TRECVID, Workshop, Gaithersburg, MD, USA
11. Barbu, A., Bridge, A., Coroian, D., Dickinson, S., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangguan, J., Siskind, J.M., Waggoner, J., Wang, S., Wei, J., Yin, Y., Zhang, Z.: Large-scale automatic labeling of video events with verbs based on event-participant interaction. In: arXiv:1204.3616v1 (2012)
12. Bay H, Ess A, Tuytelaars T, van Gool L (2008) SURF: speeded up robust features. Comput Vision Image Underst 110(3):346–359
13. Beal MJ, Jojic N, Attias H (2003) A graphical model for audio-visual object tracking. IEEE Trans Pattern Anal Mach Intell 25(7):828–836
14. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In: Proceedings of International Conference on Computer Vision
15. Bobick AF (1997) Movement, activity, and action: the role of knowledge in the perception of motion. Philos Trans Royal Soc London 352:1257–1265
16. Boiman O, Shechtman E, Irani M (2008) In defense of nearest-neighbor based image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition
17. Brezeale D, Cook D (2008) Automatic video classification: a survey of the literature. IEEE Trans Syst Man Cybernet Part C 38(3):416–430
18. de Campos C, Ji Q (2011) Efficient structure learning of bayesian networks using constraints. J Mach Learn Res 12(3):663–689
19. Cao J, Zhang YD, Song YC, Chen ZN, Zhang X, Li JT (2009) MCG-WEBV: a benchmark dataset for web video analysis. Tech. rep., ICT-MCG-09-001, Institute of Computing Technology, Chinese Academy of Sciences
20. Castel C, Chaudron L, Tessier C (1996) What is going on? a high level interpretation of sequences of images. In: Proceedings of European conference on computer vision, Springer-Verlag, London, UK
21. Chang SF, He J, Jiang YG, El Khoury E, Ngo CW, Yanagawa A, Zavesky, E. (2008) Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search. In: Proceedings of NIST TRECVID, Workshop, Gaithersburg
22. Chang YL, Zeng W, Kamel I, Alonso R (1996) Integrated image and speech analysis for content-based video indexing. In: Proceedings of IEEE international conference on multimedia computing and systems, Washington, DC
23. Chen M, Xu ZE, Weinberger KQ, Sha F (2012) Marginalized stacked denoising autoencoders for domain adaptation. In: Proceedings international conference on machine learning
24. Cheng H et al (2011) Team SRI-Sarnoff's AURORA System @ TRECVID 2011. In: Proceedings of NIST TRECVID, Workshop
25. Connolly CI (2007) Learning to recognize complex actions using conditional random fields. In: Proceedings of International Conference on Advances in Visual Computing
26. Cotton CV, Ellis DPW, Loui AC (2011) Soundtrack classification by transient events. In: Proceedings of IEEE international conference acoustics, speech, signal processing, pp 473–476
27. Dalal N, Triggs B (2005) Histogram of oriented gradients for human detection. In: Proceedings of IEEE conference on computer vision and pattern recognition
28. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of IEEE conference on computer vision and, pattern recognition
29. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In Proceedings of joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance
30. Dorko G (2012) Interest point detectors local descriptors. http://lear.inrialpes.fr/people/dorko/downloads.html
31. Duan L, Xu D, Tsang IW, Luo J (2010) Visual event recognition in videos by learning from web data. In: Proceedings of IEEE conference on computer vision and, pattern recognition
32. Duchenne O, Laptev I, Sivic J, Bach F, Ponce J (2009) Automatic annotation of human actions in video. In: Proceedings of IEEE international conference on computer vision
33. Eronen A, Peltonen V, Tuomi J, Klapuri A, Fagerlund S, Sorsa T, Lorho G, Huopaniemi J (2006) Audio-based context recognition. IEEE Trans Audio Speech Lang Process 14(1):321–329
34. Everingham M, van Gool L, Williams CKI, Winn J, Zisserman A (2007) The PASCAL visual object classes challenge 2007 (VOC2007) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/ results/index.shtml
35. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE Trans Pattern Anal Mach Intell 32(9):1530–1535
36. Feng Y, Lapata M (2010) How many words is a picture worth? automatic caption generation for news images. In: Proceedings of the annual meeting of the association for computational linguistics
37. Fillmore CJ (1968) The case for case. In: Bach E, Harms R (eds), Universals in Linguistic Theory, New York, pp 1–88
38. Fiscus J et al (2011) TRECVID multimedia event detection evaluation plan. http://www.nist.gov/itl/iad/mig/upload/MED11-EvalPlan-V03-20110801a.pdf
39. Francois ARJ, Nevatia R, Hobbs J, Bolles RC (2005) Verl: an ontology framework for representing and annotating video events. IEEE Multimedia Magazine 12(4):76–86
40. Frey BJ, Jojic N (2005) A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Trans Pattern Anal Mach Intell 27(9):1392–1416
41. van Gemert JC, Veenman CJ, Smeulders AWM, Geusebroek JM (2010) Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell 32(7):1271–1283
42. Ghanem N, DeMenthon D, Doermann D, Davis L (2004) Representation and recognition of events in surveillance video using

petri nets. In: Proceedings of IEEE conference on computer vision and pattern recognition workshop
43. Granger C (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438
44. Hakeem A, Sheikh Y, Shah M (2004) Casee: a hierarchical event representation for the analysis of videos. In: Proceedings of AAAI conference
45. Herbrich R (2001) Learning Kernel classifiers: theory and algorithms. The MIT Press, Cambridge
46. Hu Y, Cao L, Lv F, Yan S, Gong Y, Huang TS (2009) Action detection in complex scenes with spatial and temporal ambiguities. In: Proceedings of IEEE international conference on computer vision
47. Huang CL, Shih HC, Chao CY (2006) Semantic analysis of soccer video using dynamic bayesian network. IEEE Trans Multimedia 8(4):749–760
48. Inoue N, Kamishima Y, Wada T, Shinoda K, Sato S (2011) TokyoTech+Canon at TRECVID 2011. In: Proceedings of NIST TRECVID Workshop
49. Intille SS, Bobick AF (2001) Recognizing planned, multiperson action. Comput Vision Image Underst 81(3):414–445
50. Ivanov YA, Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22(8):852–872
51. Jiang W, Cotton C, Chang SF, Ellis D, Loui AC (2009) Short-term audio-visual atoms for generic video concept classification. In: Proceedings of ACM international conference on multimedia
52. Jiang W, Loui AC (2011) Audio-visual grouplet: Temporal audio-visual interactions for general video concept classification. In: Proceedings of ACM international conference on multimedia
53. Jiang YG (2012) SUPER: Towards real-time event recognition in Internet videos. In: Proceedings of ACM international conference on multimedia retrieval
54. Jiang YG, Dai Q, Xue X, Liu W, Ngo CW (2012) Trajectory-based modeling of human actions with motion reference points. In: Proceedings of European conference on computer vision
55. Jiang YG, Ngo CW, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of ACM international conference on image and video retrieval
56. Jiang YG, Yang J, Ngo CW, Hauptmann AG (2010) Representations of keypoint-based semantic concept detection: a comprehensive study. IEEE Trans Multimedia 12(1):42–53
57. Jiang YG, Ye G, Chang SF, Ellis D, Loui AC (2011) Consumer video understanding: a bechmark database and an evaluation of human and machine performance. In: Proceedings of ACM international conference on multimedia retrieval
58. Jiang YG, Zeng X, Ye G, Bhattacharya S, Ellis D, Shah M, Chang SF (2010) Columbia-UCF TRECVID2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In: Proceedings of NIST TRECVID, Workshop
59. Joo SW, Chellappa R (2006) Attribute grammar-based event recognition and anomaly detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, Workshop
60. Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of IEEE conference on computer vision and pattern recognition
61. Klaser A, Marszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: Proceedings of British machine vision conference
62. Knopp J, Prasad M, Willems G, Timofte R, van Gool L (2010) Hough transform and 3D SURF for robust three dimensional classification. In: Proceedings of European conference on computer vision
63. Kojima A, Tamura T, Fukunaga K (2002) Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vision 50(2):171–184
64. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proceedings of IEEE international conference on computer vision
65. Laptev I (2005) On space-time interest points. Int J Comput Vision 64:107–123
66. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of IEEE conference on computer vision and pattern recognition
67. Lavee G, Rivlin E, Rudzsky M (2009) Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in videos. IEEE Trans Syst Man Cybernet Part C 39(5):489–504
68. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition
69. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proceedings of IEEE conference on computer vision and, pattern recognition
70. Lee K, Ellis DPW (2010) Audio-based semantic concept classification for consumer video. IEEE Trans Audio Speech Lang Process 18(6):1406–1416
71. Li W, Zhang Z, Liu Z (2008) Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Trans Circ Syst Video Technol 18(11):1499–1510
72. Lindeberg T (1998) Feature detection with automatic scale selection. Int J Comput Vision 30:79–116
73. Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: Proceedings of IEEE conference on computer vision and, pattern recognition, pp 3337–3344
74. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In: Proceedings of IEEE conference on computer vision and pattern recognition
75. Liu J, Shah M (2008) Learning human actions via information maximization. In: Proceedings of IEEE conference on computer vision and pattern recognition
76. Loui AC, Luo J, Chang SF, Ellis D, Jiang W, Kennedy L, Lee K, Yanagawa A (2007) Kodak's consumer video benchmark data set: concept definition and annotation. In: Proceedings of ACM international workshop on multimedia, information retrieval
77. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60:91–110
78. Lu L, Hanjalic A (2008) Audio keywords discovery for text-like audio content analysis and retrieval. IEEE Trans Multimedia 10(1):74–85
79. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of international joint conference on artificial intelligence
80. Lyon RF, Rehn M, Bengio S, Walters TC, Chechik G (2010) Sound retrieval and ranking using sparse auditory representations. Neural Comput 22(9):2390–2416
81. Maji S, Berg AC, Malik J (2008) Classification using intersection kernel support vector machines is efficient. In: Proceedings of IEEE conference on computer vision and pattern recognition
82. Mandel MI, Ellis DPW (2005) Song-level features and support vector machines for music classification. In: Proceedings of international society of music information retrieval conference
83. Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 18(8):837–842

84. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M (1997) The det curve in assessment of detection task performance. In: Procedings of European conference on speech communication and technology, pp 1895–1898
85. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British machine vision conference, vol 1, pp 384–393
86. MediaEval: Multimedia retrieval benchmark evaluation. http://www.multimediaeval.org
87. Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In: Proceedings of IEEE international conference on computer vision
88. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vision 60:63–86
89. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
90. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J et al (2005) A comparison of affine region detectors. Int J Comput Vision 65(1/2):43–72
91. Minami K, Akutsu A, Hamada H, Tonomura Y (1998) Video handling with music and speech detection. IEEE Multimedia Magazine 5:17–25
92. Moore D, Essa I (2001) Recognizing multitasked activities using stochastic context-free grammar. In: Proceedings of AAAI conference
93. Moosmann F, Nowak E, Jurie F (2008) Randomized clustering forests for image classification. IEEE Trans Pattern Anal Mach Intell 30(9):1632–1646
94. Morsillo N, Mann G, Pal C (2010) Youtube scale, large vocabulary video annotation, Chapter 14 in video search and mining. Springer-Verlag series on studies in computational intelligence. Springer, Berlin, pp 357–386
95. Naphade M, Smith J, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimedia Magazine 13(3):86–91
96. Natarajan P et al (2011) BBN VISER TRECVID 2011 multimedia event detection system. In: Proceedings of NIST TRECVID, Workshop
97. Natarajan P, Nevatia R (2008) Online, real-time tracking and recognition of human actions. In: Proceedings of IEEE workshop on motion and video, computing, pp 1–8
98. Natsev A, Smith JR, Hill M, Hua G, Huang B, Merler M, Xie L, Ouyang H, Zhou, M (2010) IBM Research TRECVID-2010 video copy detection and multimedia event detection system. In: Proceedings of NIST TRECVID, Workshop
99. NIST Trecvid Multimedia Event Detection (MED) task. http://www.nist.gov/itl/iad/mig/med.cfm
100. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of IEEE conference on computer vision and pattern recognition
101. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: Proceedings of European conference on computer vision
102. Oikonomopoulos A, Patras I, Pantic M (2011) Spatiotemporal localization and categorization of human actions in unsegmented image sequences. IEEE Trans Image Process 20(4):1126–1140
103. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
104. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vision 42:145–175
105. Ordonez V, Kulkarni G, Berg TL (2011) Im2Text: describing images using 1 million captioned photographs. In: Proceedings of advances in neural information processing systems
106. Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the annual meeting of the association for computational linguistics
107. Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand M (1992) Complex sounds and auditory images. In: Proceedings of international symposium on hearing, pp 429–446
108. Perronnin F, Sanchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: Proceedings of European conference on computer vision
109. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: improving particular object retrieval in large scale image databases. In: Proceedings of IEEE conference on computer vision and pattern recognition
110. Pollard C, Sag I (1994) Head-driven phrase structure grammar. Chicago University Press, Chicago
111. Poppe R (2010) Survey on vision-based human action recognition. Image Vision Comput 28(6):976–990
112. Rapantzikos K, Avrithis Y, Kollias S (2009) Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition
113. Raptis M, Soatto S (2010) Tracklet descriptors for action modeling and video analysis. In: Proceedings of European conference on computer vision
114. Rodriguez MD, Ahmed J, Shah M (2008) Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: Procedings of IEEE conference on computer vision and pattern recognition
115. Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. Int J Comput Vision 40(2):99–121
116. Russell B, Torralba A, Murphy K, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vision 77(1–3):157–173
117. Ryoo MS, Aggarwal JK (2006) Recognition of composite human activities through context-free grammar based representation. In: Proceedings pf IEEE conference on computer vision and pattern recognition
118. Sadlier DA, O'Connor NE (2005) Event detection in field sports video using audio-visual features and a support vector machine. IEEE Trans Circ Syst Video Technol 15(10):1225–1233
119. van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596
120. Satkin S, Hebert M (2010) Modeling the temporal extent of actions. In: Proceedings of European conference on computer vision
121. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of international conference on pattern recognition
122. Scovanner P, Ali S, Shah M (2007) A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of ACM international conference on multimedia
123. Shechtman E, Irani M (2007) Matching local self-similarities across images and videos. In: Proceedings lo IEEE conference on computer vision and pattern recognition
124. Shotton J, Johnson M, Cipolla R (2008) Semantic texton forests for image categorization and segmentation. In: Proceedings of IEEE conference on computer vision and pattern recognitio
125. Si Z, Pei M, Yao B, Zhu SC (2011) Unsupervised learning of event and-or grammar and semantics from video. In: Proceedings IEEE international conference on computer vision
126. Silpa-Anan C, Hartley R (2008) Optimised KD-trees for fast image descriptor matching. In: IEEE conference on computer vision and pattern recognition

127. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Proceedings of IEEE international conference on computer vision
128. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proceedings of ACM international workshop on multimedia information retrieval
129. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
130. Snoek CGM, Worring M (2008) Concept-based video retrieval. Found Trends Inf Retr 2(4):215–322
131. Starner TE (1995) Visual recognition of american sign language using hidden markov models. Ph.D. thesis
132. Sun J, Wu X, Yan S, Cheong LF, Chua TS, Li J (2009) Hierarchical spatio-temporal context modeling for action recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition
133. Sun SW, Wang YCF, Hung YL, Chang CL, Chen KC, Cheng SS, Wang HM, Liao HYM (2011) Automatic annotation of web videos. In: Proceedings of IEEE international conference on multimedia and expo
134. Tan CC, Jiang YG, Ngo CW (2011) Towards textually describing complex video contents with audio-visual concept classifiers. In: Proceedings of ACM international conference on multimedia
135. Taylor G, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: Proceedings of European conference on computer vision
136. Torresani L, Szummer M, Fitzgibbon A (2010) Efficient object category recognition using classemes. In: Proceedings of European conference on computer vision
137. Tran SD, Davis LS (2008) Event modeling and recognition using markov logic networks. In: Proceedings of European conference on computer vision
138. Tsekeridou S, Pitas I (2001) Content-based video parsing and indexing based on audio-visual interaction. IEEE Transactions on Circuits and Systems for Video Technology 11(4):522–535
139. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circ Syst Video Technol 18(11):1473–1488
140. Tuytelaars T (2010) Dense interest points. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2281–2288
141. Uemura H, Ishikawa S, Mikolajczyk K (2008) Feature tracking and motion compensation for action recognition. In: Proceedings British machine vision conference
142. Uijlings JRR, Smeulders AWM, Scha RJH (2010) Real-time visual concept classification. IEEE Trans Multimedia 12(7): 665–680
143. University of Central Florida 50 human action dataset (2010). http://server.cs.ucf.edu/~ision/data/UCF50.rar
144. Vail DL, Veloso MM, Lafferty JD (2007) Conditional random fields for activity recognition. In: Proceedings of international joint conference on autonomous agents and multiagent systems
145. Vedaldi A, Gulshan V, Varma M, Zisserman A (2009) Multiple kernels for object detection. In: Proceedings of IEEE international conference on computer vision
146. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Procedings of international conference on machine learning
147. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11(12):3371–3408
148. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In; Proceedings of IEEE conference on computer vision and pattern recognition
149. Wang F, Jiang YG, Ngo CW (2008) Video event detection using motion relativity and visual relatedness. In: Proceedings of ACM international conference on multimedia
150. Wang H, Klaser A, Schmid C, Liu CL (2011) Action recognition by dense trajectories. In: Proceedings of IEEE conference on computer vision and pattern recognition
151. Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2008) Evaluation of local spatio-temporal features for action recognition. In: Proceedings of British machine vision conference
152. Wang J, Kumar S, Chang SF (2010) Semi-supervised hashing for scalable image retrieval. In: Proceedings of IEEE conference on computer vision and pattern recognition
153. Wang L, Suter D (2007) Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In: Proceedings of IEEE conference on computer vision and pattern recognition
154. Wang Y, Mori G (2009) Max-margin hidden conditional random fields for human action recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition
155. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vision Image Underst 104(2):249–257
156. Weiss Y, Torralba A, Fergus R (2008) Spectral hashing. In: Proceedings of advances in neural information processing systems
157. White B, Yeh T, Lin J, Davis L (2009) Web-scale computer vision using mapreduce for multimedia data mining. In: Proceedings of ACM SIGKDD workshop on multimedia data mining
158. Willems G, Tuytelaars T, van Gool L (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proceedings European conference on computer vision
159. Wu S, Oreifej O, Shah M (2011) Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: Proceedings of IEEE international conference on computer vision
160. Xie L, Xu P, Chang SF, Divakaran A, Sun H (2004) Structure analysis of soccer video with domain knowledge and hidden markov models. Pattern Recognit Lett 25(7):767–775
161. Xu C, Wang J, Lu H, Zhang Y (2008) A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans Multimedia 10(3):421–436
162. Xu D, Chang SF (2008) Video event recognition using Kernel methods with multilevel temporal alignment. IEEE Trans Pattern Anal Mach Intell 30(11):1985–1997
163. Xu M, Maddage NC, Xu C, Kankanhalli M, Tian Q (2003) Creating audio keywords for event detection in soccer video. In: Proceedings IEEE international conference on multimedia and expo
164. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden markov model. In: Proceedings of IEEE conference on computer vision and pattern recognition
165. Yan R, Fleury MO, Merler M, Natsev A, Smith JR (2009) Large-scale multimedia semantic concept modeling using robust subspace bagging and mapreduce. In: Proceedings of ACM workshop on large-scale multimedia retrieval and mining
166. Yanagawa A, Hsu W, Chang SF (2006) Brief descriptions of visual features for baseline trecvid concept detectors. Columbia University, Tech. rep.
167. Yao B, Yang X, Lin L, Lee M, Zhu S (2010) I2T: Image parsing to text description. Proc IEEE 98(8):1485–1508
168. Ye G, Jhuo IH, Liu D, Jiang YG, Chang SF (2012) Joint audio-visual bi-modal codewords for video event detection. In: Proceedings of ACM international conference on multimedia retrieval
169. Ye G, Liu D, Jhuo IH, Chang SF (2012) Robust late fusion with rank minimization. In: Proceedings IEEE conference on computer vision and pattern recognition

170. Yu TH, Kim TK, Cipolla R (2010) Real-time action recognition by sptiotemoral semantic and structural forests. In: Proceedings of British machine vision conference
171. Yuan F, Prinet V, Yuan J (2010) Middle-level representation for human activities recognition: the role of spatio-temporal relationships. In: Proceedings of ECCV Workshop on human motion: understanding, modeling, capture and animation
172. Yuen J, Russell BC, Liu C, Torralba A (2009) LabelMe video: building a video database with human annotations. In: Proceedings of international conference on computer vision
173. Zhang D, Chang SF (2002) Event detection in baseball video using superimposed caption recognition. In: Proceedings of ACM international conference on multimedia
174. Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vision 73(2):213–238