Multimed Tools Appl (2011) 51:279–302
DOI 10.1007/s11042-010-0643-7

Event detection and recognition for semantic annotation of video

Lamberto Ballan · Marco Bertini · Alberto Del Bimbo · Lorenzo Seidenari · Giuseppe Serra

Published online: 10 November 2010

Abstract Research on methods for detection and recognition of events and actions in videos is receiving an increasing attention from the scientific community, because of its relevance for many applications, from semantic video indexing to intelligent video surveillance systems and advanced human-computer interaction interfaces. Event detection and recognition requires to consider the temporal aspect of video, either at the low-level with appropriate features, or at a higher-level with models and classifiers than can represent time. In this paper we survey the field of event recognition, from interest point detectors and descriptors, to event modelling techniques and knowledge management technologies. We provide an overview of the methods, categorising them according to video production methods and video domains, and according to types of events and actions that are typical of these domains.

Keywords Video annotation · Event classification · Action classification · Survey

1 Introduction

Semantic annotation of video content is a fundamental process that allows the creation of applications for semantic video database indexing, intelligent surveillance

L. Ballan · M. Bertini (✉) · A. Del Bimbo · L. Seidenari · G. Serra
Media Integration and Communication Center, University of Florence, Florence, Italy
e-mail: bertini@dsi.unifi.it

L. Ballan
e-mail: ballan@dsi.unifi.it

A. Del Bimbo
e-mail: delbimbo@dsi.unifi.it

L. Seidenari
e-mail: seidenari@dsi.unifi.it

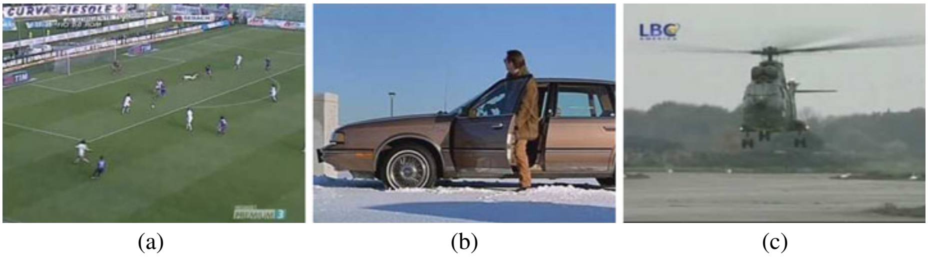
G. Serra
e-mail: serra@dsi.unifi.it

systems and advanced human-computer interaction systems. Typically videos are automatically segmented in shots and a representative keyframe of each shot is analysed to recognise the scene and the objects shown, thus treating videos like a collection of static images and losing the temporal aspect of the media.

This approach is not feasible for the recognition of events and activities, especially if we consider videos that have not been edited and do not contain shots. Recognising the presence of concepts that have a temporal component in a video sequence, if the analysis is done using simply a keyframe, is difficult [101] even for a human annotator, as shown in Fig. 1. A revision of the TRECVid 2005 ground truth annotation of 24 concepts related to events and activities has shown that 22% of the original manual annotations, performed inspecting only one keyframe per shot, were wrong [44]. An event filmed in a video is related to the temporal aspect of the video itself and to some changes in the properties of the entities and scenes represented; therefore there is need of representing and modelling time and properties' variations, using appropriate detectors, feature descriptors and models.

Several surveys on semantic video annotation have been recently presented. A review of multi-modal video indexing was presented in [94], considering entertainment and informative video domains. Multi-modal approaches for video classification have been surveyed in [19]. A survey on event detection has been presented in [56], focusing on modelling techniques; our work extends this, providing also a review of low-level features suitable for event representation, like detectors and descriptors of interest points, as well as a review of knowledge representation tools like ontologies. A survey on behaviour recognition in surveillance applications has been provided in [47], while in [81] are reported the most recent works on human action recognition. A survey of crowd analysis methods was reported in [111]. In this paper we survey methods that have been applied to different video domains, considering edited videos (i.e. videos that have been created from a collection of video material, selecting what elements to retain, delete, or combine, like movies) and unedited videos (i.e. videos that have not been processed and are simply the result of video recording, like surveillance videos). A categorisation of events and actions related to different video domains and production methods is provided, in a unified schema (see Fig. 2).

The paper is structured as follows: in the next section are briefly reviewed approaches for semantic video annotation; in Section 3 we propose a classification of events and activities; the state-of-the-art of features suitable for event and action



(a) (b) (c)

Fig. 1 Keyframe-based video event recognition. (**a**) Is it *shot-on-goal* or *placed-kick*? (**b**) Is the person *entering* or *exiting* in/from the car? (**c**) Is the aircraft *landing* or *taking-off*?

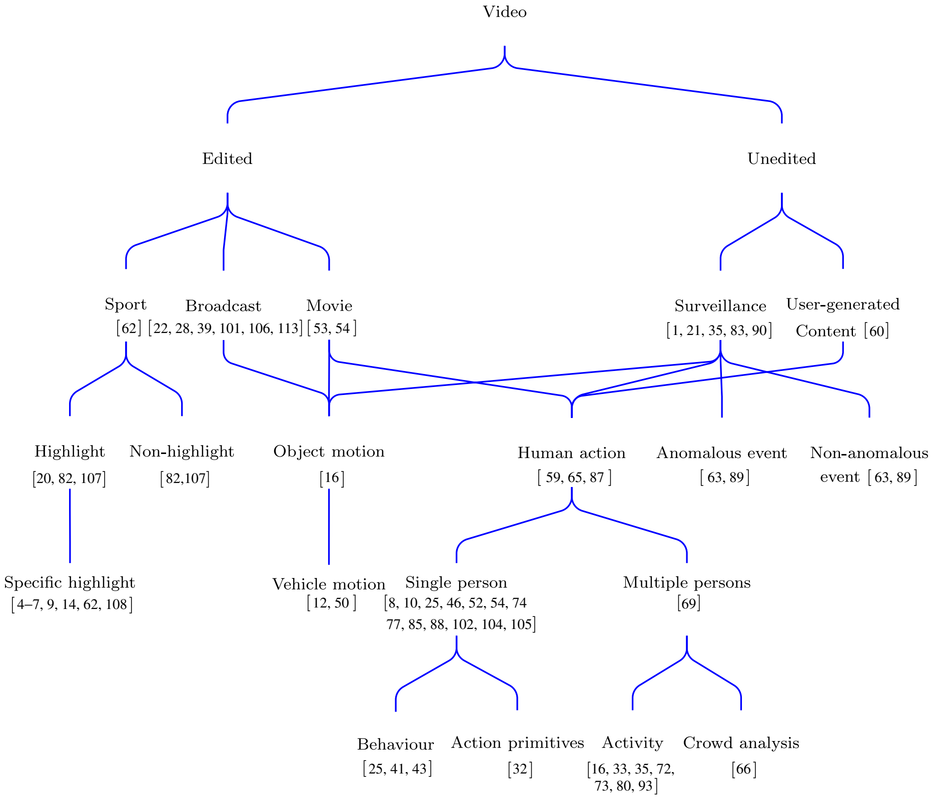

Video
Edited
Unedited
Sport [62]
Broadcast [22, 28, 39, 101, 106, 113]
Movie [53, 54]
Surveillance [1, 21, 35, 83, 90]
User-generated Content [60]
Highlight [20, 82, 107]
Non-highlight [82,107]
Object motion [16]
Human action [59, 65, 87]
Anomalous event [63, 89]
Non-anomalous event [63, 89]
Specific highlight [4–7, 9, 14, 62, 108]
Vehicle motion [12, 50]
Single person [8, 10, 25, 46, 52, 54, 74 77, 85, 88, 102, 104, 105]
Multiple persons [69]
Behaviour [25, 41, 43]
Action primitives [32]
Activity [16, 33, 35, 72, 73, 80, 93]
Crowd analysis [66]

Fig. 2 From *top to bottom*: overview of types of video production, video domains and events, with references to methods proposed in literature to deal with them

representation are presented in Section 4; models and classifiers are discussed in Section 5, while ontological representations of domain knowledge are surveyed in Section 6. Finally, conclusions are drawn in Section 7.

2 Semantic video annotation

The problem of semantic video annotation is strictly related to the problem of generic visual categorisation, like classification of objects or scenes, rather than that of recognising a specific class of objects. Recently it has been shown that part-based approaches are effective methods for scene and object recognition [31, 91, 110, 112] due to the fact that they can cope with partial occlusions, clutter and geometrical transformations. Many approaches have been presented, but a common idea is to model a complex object or a scene by a collection of local interest points. Each of these local features describes a small region around the interest point therefore achieving robustness against occlusion and clutter. To deal effectively with changes of viewing conditions the features should be invariant to geometrical transformations such as translation, rotation, scaling and also affine transformations. SIFT [61] and

SURF [13] features have become the de facto standards, because of their good performance and (relatively) low computational cost. In this field, a solution that recently has become very popular is the Bag-of-Words (BoW) approach. It has been originally proposed for information retrieval, where it is used for document categorisation in a text corpus, where each document is represented by its word frequency. In the visual domain, an image or a frame of a video is the visual analogue of a document and it can be represented by a bag of quantised invariant local descriptors, called *visual-words*. The main reason for the success of this approach is that it provides methods that are sufficiently generic to cope with many object types simultaneously. The efficacy of the BoW approach is demonstrated also by the large number of systems based on this approach that participate in the PASCAL VOC and TRECVid [92] challenges.

More recently, the problem of the detection and recognition of events and activities is getting a larger attention, also within the TRECVid evaluation: the high-level concept detection task of TRECVid 2009 [78] considered the problem of event detection, with 7 out of 20 high-level concepts to be detected that were related to events and actions [22]. The most recent approaches proposed in this task have started to cope with the problem of representing videos considering the temporal aspects of it, analysing more than one keyframe per shot and introducing some representation of the context [78, 109]. Since 2008 a new dataset of airport surveillance videos, to be used in a event detection task, has been added to the TRECVid evaluation campaign; the dataset focuses mostly on crowd/group actions (e.g. people meeting), human gestures (e.g. person running) and human activities (e.g. putting an object somewhere).

3 Events and actions

We refer to events as concepts with a dynamic component; an *event* is "something happening at a given time and in a given location". In the video analysis community the event recognition task has never been tackled by proposing a generic automatic annotation tool and the proposed approaches are usually domain dependent. Video domains considered in this survey are broadcast news, sports, movies, video-surveillance and user generated content. Videos in the broadcast news, sports and movies are usually professionally edited while video-surveillance footage and user generated content are usually unedited. This editing process adds a structure [94] which can be exploited in the event modelling as explained in Sections 5 and 6. Automatic annotation systems are built so as to detect events of interest. Therefore we can firstly split events in *interesting* and *non-interesting*; in the case of video-surveillance interesting events can be specific events such as "people entering a prohibited area", "person fighting" or "person damaging public property", and so on; sometimes defining a-priori these dangerous situations can be cumbersome and, of course, there is the risk of the non exhaustivity of the system; therefore it can be useful to detect *anomalous* vs. *non-anomalous* (i.e. normal) events [63, 89]. In this case an event is considered interesting without looking at its specific content but considering how likely is given a known (learnt) statistics of the regular events. Also in the sport domain the detection of rare events is of interest, but systems need to detect events with a specific content (typically called *highlights*, [14]) such as "scoring goal”, “slam dunk”, “ace serve”, etc. Most of the domains in which video-analysis is performed involve the analysis of human motion (sports, video-surveillance, movies). Events originated by human motion can be of different complexity, involving one or more subjects and either lasting few seconds or happening in longer timeframes. *Actions* are short task oriented body movements such as “waving a hand”, or “drinking from a bottle”. Some actions are atomic but often actions of interest have a cyclic nature such as “walking” or “running”; in this case detectors are built to recognise a single phase of it. Actions can be further decomposed in *action primitives*, for example the action of running involves the movement of several body limbs [32]. This kind of human events are usually recognised using low-level features, which are able to concisely describe such primitives, and using per-action detectors trained on exemplar sequences. A main difficulty in the recognition of human actions is the high intra-class variance; this is mainly due to variation in the appearance, posture and behaviour (i.e. “the way in which one acts or conducts oneself”) of the “actor”; *behaviour* can thus be exploited as a biometric cue [43].

Events involving multiple people or happening in longer timeframes can be referred as *activities* [81]. Activity analysis requires higher level representations usually built with action detectors and reasoning engines. Events can be defined activities as long as there is not excessive inter-person occlusion and thus a system is able to analyse each individual motion (typically in sequences with two to ten people). In case of presence of a large amount of people, the task is defined as *crowd analysis* [111]: persons are no more considered as individuals but the global motion of a crowd is modelled [66]. In this case the detection of anomalous events is prominent because of its applicability to surveillance scenarios and because of the intrinsic difficulty of precisely defining crowd behaviours. *Human actions* are extremely useful in defining the video semantics in the domains of movies and user generated content. In both domains the analysis techniques are similar and challenges arise mainly from the high intra-class variance. Contextual information such a static features or scene classifiers may improve event recognition performance [39, 60, 65].

In the broadcast news domain several events of interest do not involve people; moreover some of them do, but more information can be obtained from contextual cues; as an example visual cues of smoke and fire, together with a detection of a urban scene can identify a riot. Also in the sport domain contextual information and its temporal evolution contain most of the information, thus no human motion analysis is usually performed to detect interesting events. Events may also relate to the motion of an object such as a vehicle, in this case we refer to *object motion* and *vehicle motion* events which are of interest in the broadcast [12] and in the video-surveillance [50] domains.

Figure 2 shows an overview of types of video production, video domains and events, and the methods proposed in literature that can recognise them.

4 Features for actions and events

Recognition of events in video streams depends on the ability of a system to build a discriminative model which has to generalise with respect to unseen data. Such generalisation is usually obtained by feeding state-of-the art statistical classifiers with an adequate amount of data. We believe that the key to solve this issue is the use of sufficiently invariant and robust image descriptors. While tackling a problem such as single-object recognition (i.e. find instances of “this object” in a given collection of images or videos) image descriptors are required to yield geometric and photometric invariance in order to match object instances across different images, possibly acquired with diverse sensors in different lighting environment and in presence of clutter and occlusions. An elegant way of dealing with clutter, occlusion and viewpoint change is the use of region descriptors [61, 67]; image regions can be normalised [68] to obtain invariance to deformations due to viewpoint change, other normalisation can be applied to obtain rotation and partial photometric invariance [61].

This kind of description has been extended in the object and scene categorisation scenario exploiting the bag-of-words framework [91]. Through the use of an intermediate description, the codebook, images are compactly represented. The codebook is usually obtained with a vector quantisation procedure exploiting some clustering algorithm such as k-means. This intermediate description allows both fast data access, by building an inverted index [75, 91], and generalisation over category of objects by representing each instance as a composition of common parts [31]. As in the textual counterpart the bag of visual words does not retain any structural information: by using this representation we actually do not care where regions occur in an image. As this comes with some advantages like robustness to occlusions and generalisation over different object and scenes layouts, there is also a big disadvantage in discarding completely image structure, since this actually removes all spatial information. A local visual words spatial layout description [84] can recover some image structure without loss of generalisation power. A global approach has been proposed by Lazebnik et al. [57]; in their work structure is added in a multi-resolution fashion by matching spatial pyramids obtained by subsequently partitioning the image and computing bag-of-words representations for each of the sub-image partition.

Given the success of bag of keypoints representations in static concept classification, efforts have been made to introduce this framework in event categorisation. The first attempt in video annotation has been made by Zhou et al. [113], describing a video as a bag of SIFT keypoints. Since keypoints are considered without any spatial or temporal location (neither at the frame level) it is possible to obtain meaningful correspondences between varying length shots and shots in which similar scenes occur in possibly different order. Again, the structure is lost but this allows a robust matching procedure. Anyway temporal structure of videos carries rich information which has to be considered in order to attain satisfactory video event retrieval results. This information can be recovered using sequence kernels, as reviewed in Section 5. A different temporal information lies at a finer grained level and can be captured directly using local features. This is the case of gestures, human actions and, to some extent, human activities. Since gestures and actions are usually composed of *action primitives*, which occur in a short span of time and involve limb movements, their nature is optimally described by a local representation.

As in static keypoint extraction frameworks, the approach consists of two stages, detection and description. The detection stage aims at producing a set of “informative regions” for a sequence of frames, while the goals of the description stage are to gain invariance with respect to several region transformations caused by the image formation process, and to obtain a feature representation that enables matching through some efficiently computable metric.

4.1 Detectors

Space-time interest points located by detectors should contain information on the objects and their motion in the world. Detectors are thus functions computed over the image plane and over time that present higher values in presence of local structures undergoing non-constant motion. These structures in the image should correspond to an object part that is moving in the world. Since they deal with dynamic content they need to be robust to motion generated by camera movements; these noisy detections have to be filtered without damaging detector ability to extract interesting image structures.

Local dynamic representations have been mostly derived directly from their static counterparts [52, 77, 102, 104] while the approaches presented in [22, 25] are explicitly designed for space-time features. Laptev extended Harris corners keypoints to the space-time domain [52]; space-time corners are corner-like structures undergoing an inversion of motion. Wong et al. employed a difference-of-Gaussian operator on space-time volumes, after a preprocessing with non-negative matrix factorisation, in order to exploit the global video structure. Willems extended the SURF [13] detector using box filters and integral videos in order to obtain almost real time feature extraction; finally, the saliency measure originally proposed by Kadir and Brady [42] have been extended by Oikonomopoulos et al. [77]. The detector proposed by Dollár et al. [25] separates the operator which process the volume in space and time; the spatial dimension is filtered with a Gaussian kernel while the temporal dimension is processed by Gabor filters in order to detect periodic motion. A similar approach, specifically designed for the spatio-temporal domain, has been proposed by Chen et al. [22], which exploits a combination of optical flow based detectors with the difference of Gaussian detector used by SIFT.

Region scale can be selected by the algorithm [52, 102, 104] both in space and time or may simply be a parameter of it [25, 54]; moreover scale for space and time can be fixed as in [25] or a dense sampling can be performed to enrich the representation [8, 54]. Figure 3 shows an example of the response of the detectors presented in [8], applied to the video surveillance domain. All the above approaches model the detector as an analytic function of the frames and scales, some other approaches instead rely on learning how to perform the detection using neural networks [45] or extending boosting and Haar features used for object detection [99]. Kienzle et al. trained a feed-forward neural network using, as a dataset, human eye fixations recorded with an headmounted tracker during the vision of a movie.

Recent detectors and approaches lean toward a denser feature sampling, since in the categorisation task a denser feature sampling yields a better performance [76]. State-of-the art image classifiers are, by now, performing feature sampling over regular multi-scale overlapped grids. This kind of approach is probably still too computational expensive to be performed on a sequence composed of hundred of frames. Finally, to the end of extracting as much information as possible, multiple feature detectors, either static or dynamic, have been used in conjunction [60, 65, 69].

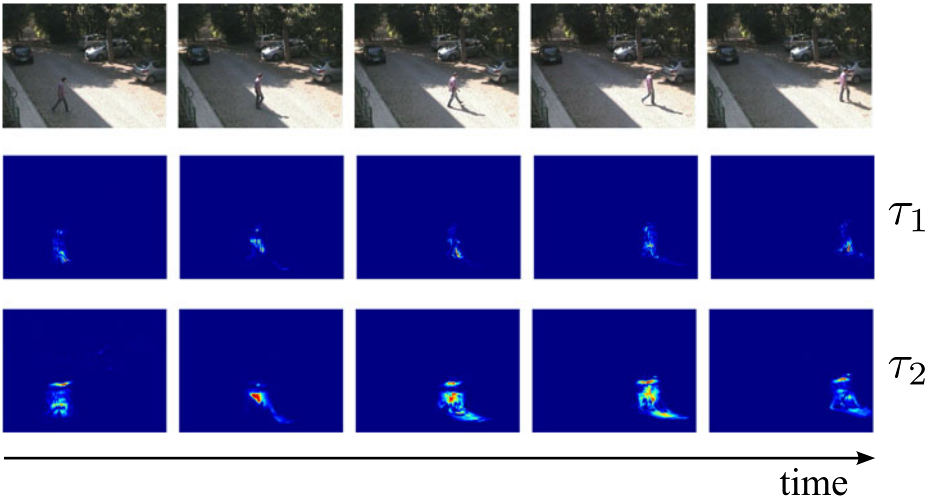
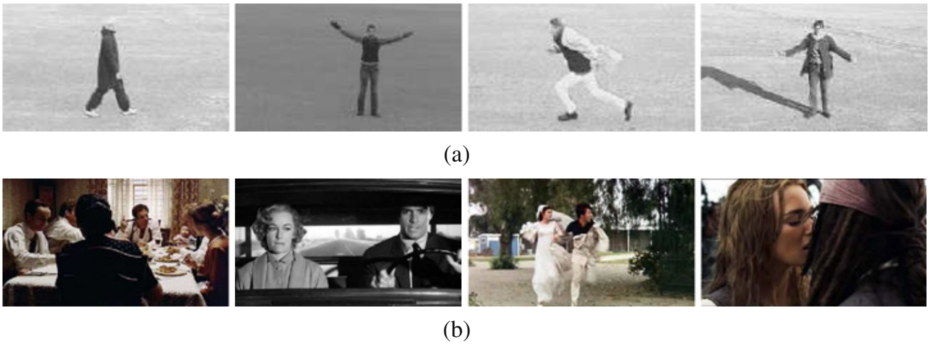


Fig. 3 Spatio-temporal interest point detector [8] running at different temporal scales (*blue* low response, *red* high response); *first row*: original video frames, *second row* detector response at temporal scale τ_1 (mostly due to the limbs), *third row*: detector response temporal scale τ_2 (mostly due to the torso), with $\tau_1 < \tau_2$. Frames taken from the ViSOR video repository [98]

4.2 Descriptors

The regions extracted by detectors need to be represented compactly. Descriptors are usually computed using a common pipeline as outlined in [103] for static features and, partially, in [51] for dynamic ones: preprocessing, non-linear transformation, pooling and normalisation. The preprocessing stage is usually a smoothing operation performed using a 3-dimensional Gaussian kernel [46, 52]. In order to obtain more robust descriptors a region normalisation can be applied [52]; the normalisation procedure proposed by Laptev attempt to obtain camera-motion invariant regions in order to increase the matching procedure reliability. Regions are transformed by computing an image measurement; typical choices are: normalised brightness [25], image gradients [52], spatio-temporal gradients [8, 25, 46, 88] and optical flow [8, 25, 52]. Gradients are used to provide photometric invariance, 3-dimensional gradients are capable of representing appearance and motion concisely. Optical flow descriptors can offer very informative low dimensional representations in case of smooth motion patterns, but in presence of noise the performance may degrade. Even if both carry motion information these two descriptions have been found to be complementary [8] and the fusion is beneficial for recognition. After computing this region transformation, the descriptor size is still very high dimensional and there is no invariance to small deformations (due for example to viewpoint change). Typically either global [25, 51] or local [8, 46, 88] histograms of gradient/optical flow orientation are computed. The use of local statistics contribute to obtain invariance to little viewpoint changes. A simpler approach is to apply PCA to the concatenated brightness, gradient or optical flow values [25, 51]. A different technique is to compute higher order derivatives of image intensity values [52]. Finally, following the approach of SIFT a descriptor normalisation and clipping can be applied to



(a)

(b)

Fig. 4 Sample frames from actions in KTH **a** and Hollywood **b** datasets

obtain robustness w.r.t. contrast change [46]. As shown in [103], for static feature descriptors, parameters can be learnt instead of "handcrafted"; Marszalek et al. performed such an optimisation by training on datasets [65]. This technique shows an improvement over the handcrafted values but it is also shows sensitivity to data: descriptors trained over Hollywood movies[1] dataset does not perform as well on videos of the KTH dataset[2] and vice-versa. Figure 4 shows sample frames of these two datasets.

4.3 Action representation

Actions can be represented as a collection of space-time pixel neighbourhoods descriptors. Statistical classification frameworks require an instance-to-instance or an instance-to-class matching procedure. Local feature matching can be done using simple metrics such as the Euclidean distance and exploiting [61] nearest neighbour distances to remove outliers. This technique is highly effective in the single-object recognition task but can deliver poor performance when generalisation power is needed as in a category recognition problem. As in object category recognition the intermediate codebook representation can offer together generalisation power and dimensionality reduction; in fact features which are often high dimensional (200+) are replaced with a code corresponding to a visual word in the dictionary. As stated previously bag-of-words representations completely lack any notion of the global features layout or their correlations. In action representation the visual words are often associated with an action primitive such as "raising an arm" or "extending a leg forward" and their spatio-temporal dependence is a strong cue. These relations can be modelled in the codebook formation [59, 88] or encoded in the final action representation [69, 74, 85, 105]. Scovanner et al. [88] have grouped co-occurring visual words to capture spatio-temporal feature correlations. Liu et al. have acted similarly on the dictionary by iteratively grouping visual words that maximise the mutual information. Niebles et al. [74] and Wong et al. [105] exploited graphical models to introduce a structural representation of the human action by modelling

[1] http://www.irisa.fr/vista/actions/

[2] http://www.nada.kth.se/cvap/actions/

relations among body parts and their motion. Savarese et al. [85] augmented the action descriptor by computing visual words spatio-temporal correlograms instead of a flat word-count. Finally Mikolajczyk and Uemura [69] exploited vocabulary forest together with a star-shape model of the human body to allow localisation together with recognition. All these structural representations deal with relations between the feature themselves and are suitable in the analysis of isolated actions or behaviours. In the case of unconstrained scenarios, global layout representation can be a better choice [29, 53, 54]. The main advantage is their reduced computational cost. Moreover their coarse description can deal better with a higher intra-class variation. These approaches split the video volume with a coarse spatio-temporal grid, which can have a uniform [29, 53] or non-uniform layout [54], and by binning features in space and time, position dependent feature statistics is computed.

5 Classification of composite events

Events that are characterised by complex or composite evolution are often modelled by using a mid-level representation of the particular domain which eases the event recognition. Therefore many works try to build classifiers that are able to characterise the evolution and the interaction of particular visual features. These kinds of representations are often used in specific domains (for example in sports videos), where it is easier to define "in advance" the relations among visual features. Several different techniques have been proposed in the literature for this purpose: simple heuristic rules, finite state machines, statistical models (such as HMM or Bayesian networks) and kernel methods.

5.1 Heuristic rules and finite state machines

Several works in the sports video domain apply heuristics or rule-based approaches to automatically recognise simple events. An example is given by Xu et al. [107] in which recognition of play/break events of soccer videos is performed using classification of simple and mutually exclusive events (obtained by using a simple rule-based approach). Their method is composed by two steps; in the first step they classify each sample frame into global, zoom-in and close-up views using an unique domain-specific feature, grass-area-ratio. Then heuristic rules are used in processing the sequence of views, and obtain play/break status of the game.

More complex events can be recognised using Finite State Machines (FSMs). The knowledge of the domain is encoded into a set of FSMs and each of them is able to represent a particular video event. This approach was initially proposed by Assfalg et al. in [5] to detect the principal soccer highlights, such as shot on goal, placed kick, forward launch and turnover, from a few visual cues, such as playground position, speed and camera direction, etc. The idea of applying FSMs to model highlights and events has been recently followed also in [7]; scored goal, foul and generic play scenes in soccer videos have been modeled using four types of views (e.g. in-field, slow motion, etc.) for the states of the FSMs and transitions are determined by some audio-visual events such as the appearance of a caption or the whistle of the referee. Experiments have been performed using a set of manually annotated views and audio-visual events.

5.2 Markovian models

Visual events that evolve in a predictable manner are suitable for a Markovian modelling, and thus they can be detected by HMMs. Sports videos, in particular those that have a specific structure due to the rules like baseball and tennis, have been analysed using HMMs for event classification. If the events always move forward then a left-to-right model may be more suitable; in other cases, if the meaning of the states is not tangible it is better to choose a model with a sufficient number of states. A fully connected (ergodic) model is more suited for unstructured events. The feature set needs to capture the essence of the event, and features have to be chosen depending on the events being modelled. In general the steps that have to be followed when using HMMs for event classification/recognition [38] is to check if a "grammar" of the events is identifiable: this helps to identify if HMMs can model events directly or if the states within the HMM model the events. An appropriate choice of model topology, e.g. left-to-right or fully connected, has to be done. Then features have to be chosen according to the events to be modelled. Enough training data, representative of the range of manifestations of the events, has to be selected, increasing its size in case of ergodic models. In general a significant effort is required to train a HMM system, and ergodic models require more training data than left-to-right models. In [18] is noted that the conventional HMM training approaches, based on maximum likelihood such as the Baum-Welch algorithm, often produce models that are both under-fit (failing to capture the hidden structure of the signal) and over-fit (with many parameters that model noise and signal bias), thus leading to both poor predictive power and small generalisation.

A number of approaches that use HMM have been proposed to analyse sports videos, since the events that are typical for this domain are very well suited for this approach. It has to be noted that reliable event classification can be achieved if events have been accurately segmented and delimited. Classification of three placed kicks events (free, corner and penalty kick) using HMMs has been proposed by Assfalg et al. in [4], using a 3-state left-to-right model for each highlight, based on the consideration that the states correspond well to the evolution of the highlights in term of characteristic content. The features used are the framing term (e.g. *close-up*), camera pan and tilt (quantised in five and two levels). Similar approaches for event detection in news videos have been applied also at a higher semantic level, using the scores provided by concept detectors as synthetic frame representations or exploiting some pre-defined relationships between concepts. For example, Ebadollahi et al. [28] proposed to treat each frame in a video as an observation, applying then HMM to model the temporal evolution of an event. In [108] multi-layer HMMs (called SG-HMM) have been proposed by Xu et al. for basket and volleyball. Each layer represents a different semantic layer, and low-level features (horizontal, vertical and radial motion and acceleration cues) are fed to the bottom layer to generate hypothesis of basic events, the upper layer gets the results of the below HMMs and each state corresponds to an HMM; this requires to treat differently these HMM: the observation probability distribution is taken from the likelihood of the basic HMMs. Fully connected HMMs, with six states, are used to model all the basic events in both sports. The Basket SG-HMM has two layers: one for sub-shot classification and the upper layer for shot classification in 16 events. The Volley SG-HMM has three layers: shots are classified in the two bottom layers, and the intermediate layer accounts for shots relationships; this allows to classify 14 events that cannot be recognised within a shot.

5.3 Bayesian networks

Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode conditional independencies between the variables. Nodes can represent any kind of variable, be it a measured parameter, a latent variable or a hypothesis. Bayesian networks can represent and solve decision problems under uncertainty. They are not restricted to representing random variables, which represents another "Bayesian" aspect of a Bayesian network. Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (such as for example speech signals or protein sequences) are called Dynamic Bayesian Networks (DBNs). Dynamic Bayesian Networks are directed graphical models of stochastic processes. They generalise hidden Markov models (HMMs). In fact a HMM has one discrete hidden node and one discrete or continuous observed node per slice. In particular a Hidden Markov Model consists of a set of discrete states, state-to-state transition probabilities, prior probabilities for the first state and output probabilities for each state.

In [62] Bayesian Networks are used to recognise frame and clip classes (close-up, playfield centre and goal areas, medium views). In order to identify shot-on-goals the proposed system groups the clips that are preceding and following the clips classified as showing the goal areas. If a certain pattern of clips is found, and the values of a feature that corresponds to the position of the field end line follow a certain pattern, then a shot-on-goal is determined to be present. In [20] DBNs are used by Chao et al. to model the contextual information provided by the timeline. It is argued that HMMs are not expressive enough when using a signal that has both temporal and spatial information; moreover, DBNs allow a set of random variables instead of only one hidden state node at each time instance: this stems from the fact that HMMs are a special case of DBNs. In [20] five events are defined and are modeled considering five types of primitive scenes such as close-ups, medium views, etc. Medium level visual features such as playfield lines are used as observable features. Since all the states of the DBN are observable in the training stage it is required to learn the initial and transition probabilities among the scenes in each event separately. In the inference stage the DBN finds the most plausible interpretation for an observation sequence of features.

5.4 Kernel methods

Kernel methods are a class of algorithms for pattern analysis, whose best known element is the Support Vector Machine (SVM), a group of supervised learning methods that can be applied to classification problems. These methods map the input data into a high dimensional feature space, by doing a non-linear transformation using suitably chosen basis functions (kernel). This is known as the "kernel trick". The linear model in the feature space corresponds to a non-linear model in the input space. The kernel contains all of the information about the relative positions of the inputs in the feature space; the actual learning algorithm is based only on the kernel function and can thus be carried out without explicit use of the feature space. Since there is no need to evaluate the feature map in the high dimensional feature space, the kernel function represents a computational shortcut.

An approach that uses SVM with RBF kernel to classify sequences that contain interesting and non-interesting events was proposed in [82], showing an application to field sports such as soccer, hockey and rugby. Each shot is represented using five values, one for each feature used (e.g. speech-band audio activity, motion activity, etc.), and the maximum value of each feature is selected as representative value for the whole shot. In this way the temporal extent and the dynamics of the event are not considered or exploited. Authors note that a classification scheme such as HMM may be more appropriate if continuous knowledge of past and present states is desired. In [39] was proposed the use of SVM models for a set of motion features, computed from MPEG motion vectors, and static features, followed by a late fusion strategy to aggregate results at the decision level.

As briefly discussed in Section 4, many methods proposed recently extend the traditional BoW approach. In fact, the application of this part-based approach to event classification has shown some drawbacks with respect to the traditional image categorisation task. The main problem is that it does not take into account temporal relations between consecutive frames, and thus event classification suffers from the incomplete dynamic representation. Recently methods have been proposed to consider temporal information of static part-based representations of video frames. Xu and Chang [106] proposed to apply Earth Mover's Distance (EMD) and Temporally Aligned Pyramid Matching (TAPM) for measuring video similarity; EMD distance is incorporated in a SVM framework for event detection in news videos. In [101], BoW is extended constructing relative motion histograms between visual words (ERMH-BoW) in order to employ motion relativity and visual relatedness. Zhou et al. [113] presented a SIFT-Bag based generative-to-discriminative framework for video event detection, providing improvements on the best results of [106] on the same TRECVid 2005 corpus. They proposed to describe video clips as a bag of SIFT descriptors by modeling their distribution with a Gaussian Mixture Model (GMM); in the discriminative stage, specialised GMMs are built for each clip and video event classification is performed. Ballan et al. [10] modelled events as a sequence composed of histograms of visual features, computed from each frame using the traditional bag-of-words (see Fig. 5). The sequences are treated as strings where each histogram is considered as a character. Event classification of these sequences of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a string kernel that uses the Needlemann–Wunsch edit distance. Hidden Markov Model Support Vector Machine (SVMHMM), which is an extension

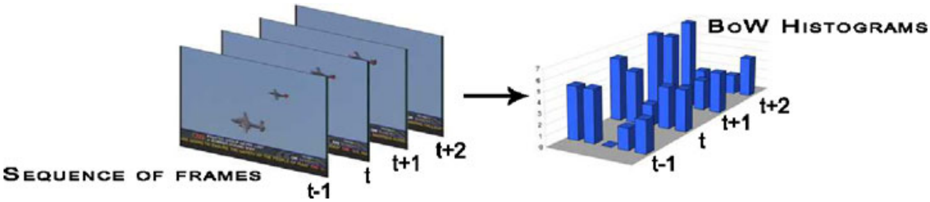


Fig. 5 Shots are represented as a sequence of BoW histogram; Events are so described by concatenation of histograms of variable size, depending on the clip length. Example taken from [10]

of the SVM classifier for sequence classification, has been used in [41] to classify the behaviour of caged mice.

6 Ontologies

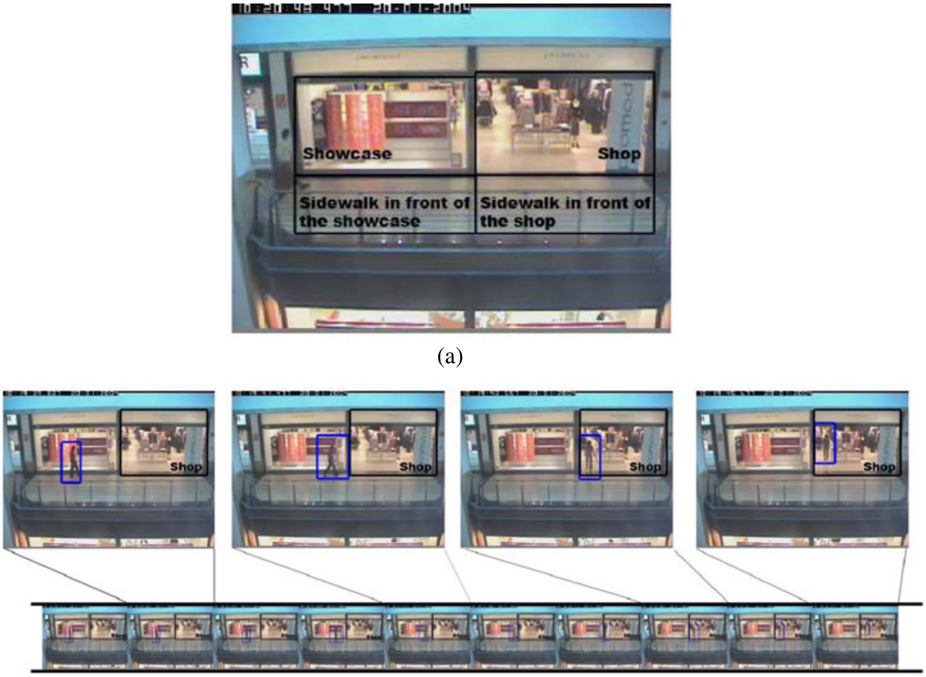
In many video content-based applications there is need of methodologies for knowledge representation and reasoning, to analyse the context of an action in order to infer an activity. This has led to an increasing convergence of research in the fields of video analysis and knowledge management. This knowledge can include heterogeneous information such as video data, features, results of video analysis algorithms or user comments. Logical-based methods for activity recognition have been proposed, to represent domain knowledge and model each event. In these approaches an event is generally specified as a set of logical rules that allow to recognise them by using logical inference techniques, such as resolution or abduction [3, 26, 79, 90]. In particular, Shet et al. [90] proposed a framework that combines computer vision algorithms with logic programming to represent and recognise activities in a parking lot in the domain of video surveillance. Lavee et al. [55] have proposed the use of Petri-Nets, and provided a methodology on how to transform ontology definitions in a Petri-Net formalism. Artikis et al. [3] and Paschke et al. [79] presented two different activity recognition systems based both on a logic programming implementation of an Event Calculus dialect [49]. The Event Calculus is a set of first-order predicate calculus, including temporal formalism, for representing and reasoning about events and their effects. These approaches do not consider the problems of noise or missing observations, that always exist in real world applications. To cope with these issues, some extensions to logic approaches have been presented. Tran et al. [95] described a domain knowledge as first-order logic production rules with associated weights to indicate their confidence. Probabilistic inference is performed using Markov-logic networks. While logic-based methods are an interesting way of incorporating domain knowledge, they are limited in their utility to specific settings for which they have been designed. Hence, there is need of a standardised and shareable representation of activity definitions.

Recently, ontologies have been regarded as the appropriate tool for domain knowledge representation because of several advantages. Their most important property is that they provide a formal framework for supporting explicit, shareable, machine-processable semantics definition of domain knowledge, and they enable the derivation of implicit knowledge through automated inference. In particular, an ontology is a formal specification of a shared conceptualisation of a domain of interest [36] and form an important part of the emerging semantic web, in which ontologies allow to organise contents through formal semantics. Ontology Web Language (OWL) and Semantic Web Rule Language (SWRL) have been proposed by the World Wide Web Consortium (W3C) as language standards for representing ontologies and rules, respectively. SPARQL Protocol and RDF Query Language (SPARQL) has been approved as W3C recommendation as query language for the Semantic Web technologies. An overview of such languages is presented in [70]. These languages enable autonomic agents to reason about Web content and to carry out more intelligent tasks on behalf of the user. Thus, ontologies are suitable for expressing video content semantics.

For these reasons, many researches have exploited ontologies to perform semantic annotation and retrieval from video digital libraries [48]. Ontologies that can be used for semantic annotation of videos are those defined by the Dublin Core Metadata Initiative [27], TV Anytime [97]—they have defined standardised metadata vocabularies—and the LSCOM initiative [71]—that has created a specialised vocabulary for news video. Other ontologies provide structural and content-based description of multimedia data, similarly to the MPEG-7 standard [2, 34, 96]. Other approaches have directly included in the ontology an explicit representation of the visual knowledge [17, 64]. Dasiopoulou et al. [23] have included in the ontology instances of visual objects. They have used as descriptors qualitative attributes of perceptual properties like colour homogeneity, low-level perceptual features like components distribution, and spatial relations. Semantic concepts have been derived from colour clustering and reasoning. In the attempt of having richer annotations, other authors have explored the usage of reasoning over multimedia ontologies. In this case spatial relationships between concept occurrences are analysed so as to distinguish between scenes and provide more precise and comprehensive descriptions. Hollink et al. [40] defined a set of rules in SWRL to perform semi-automatic annotation of images. Jain et al. [58] have employed a two-level ontology of artistic concepts that includes visual concepts such as colour and brushwork in the first level, and artist name, painting style and art period for the high-level concepts of the second level. A transductive inference framework has been used to annotate and disambiguate high-level concepts. In Staab et al. [24] automatically segmented image regions are modeled through low-level visual descriptors and associated to semantic concepts using manually labelled regions as training set. Context information is exploited to reduce annotation ambiguities. The labelled images are transformed into a constraint satisfaction problem (CSP), that can be solved using constraint reasoning techniques.

Several authors have exploited ontologies for event recognition. These methods have to deal with two issues: how to represent the entities and events of the considered domain in the ontology, and how to use the ontology for improving the video event analysis results. For solving the first issue, researchers have proposed ontologies to describe several domains, e.g. for visual surveillance analysis. In particular, Hakeen and Shah [37] have defined a meeting ontology that is determined by the knowledge base of various meeting sequences. Chen et al. [21] proposed an ontology for analysing social interaction of the patients with one another and their caregivers in a nursing home, and Georis et al. [35] for describing bank attack scenarios. Akdemir et al. [1] drew on general ontology design principles and adapted them to the specific domains of human activity, bank and airport tarmac surveillance. Moreover, a special formal language to define ontologies of events, that uses Allen's logic to model the relations between the temporal intervals of elementary concepts so as to be able to assess complex events in video surveillance has been proposed by Francois et al. [33, 73]. More recently, Scherp et al. [86] defined a formal model of events that allows interchange of event information between different event-based systems, causal relationships between events, and interpretations of the same event by different humans. A more generic approach has been followed in [80], where a verb ontology has been proposed to better describe the relations between events, following Fellbaum's verb entailments [30]. This ontology is used to classify events that may help the comprehension of other events (e.g. when an event is a precondition of another one). The outcomes of event classification are then used to create hyperlinks between video events using MPEG-7 video annotations, to create a hypervideo.

Solutions for the second issue have also been explored. Neumann and Möller [72] have proposed a framework for scene and event interpretation using Description Logic reasoning techniques over "aggregates"; these are composed of multiple parts and constrained by temporal and spatial relations to represent high-level concepts, such as objects configurations, events and episodes. Another solution was presented by Bertini et al. in [15], using generic and domain specific descriptors, identifying visual prototypes as representative elements of visual concepts and introducing mechanisms for their updating, as new instances of visual concepts are added to the ontology; the prototypes are used to classify events and objects observed in video sequences. Bai et al. [6] defined a soccer ontology and applied temporal reasoning with temporal description logic to perform event annotation in soccer videos. Snidaro et al. [93] addressed the problem of representing complex events in the context of security applications. They described a complex event as a composition of simple events, thus fusing together different information, through the use of the SWRL language. SWRL rules have been also employed to derive complex events in soccer domain [9]. In [83] the authors proposed an ontology that integrates two kinds of knowledge information: the scene and the system. Scene knowledge is described in


Showcase
Shop
Sidewalk in front of the showcase
Sidewalk in front of the shop
Shop
Shop
Shop
Shop

(a)

(b)

Fig. 6 **a** CAVIAR Surveillance video dataset: view of the mall shop areas. **b** Example of person detector and tracking in a video sequence. Example taken from [16].

Rule: PersonEntersShop

$Person(?p) \land Clip(?c) \land PersonIsInFrontShop(?p, ?g1) \land PersonIsInShop(?p, ?g2) \land$
$Temporal : notOverlaps(?g2, ?g1) \land Temporal : notBefore(?g2, ?g1) \land$
$Temporal : notMetBy(?g2, ?g1) \land HasTemporalPeriod(?c, ?g3) \land$
$Temporal : contains(?g3, ?g1) \land Temporal : contains(?g3, ?g2) \rightarrow PersonEntersShop(?p, ?c)$

Fig. 7 Rule for human action recognition, obtained using FOILS [16]. Variables are indicated using the standard convention of prefixing them with a question mark

terms of objects and relations between them. System knowledge is used to determine the best configuration of the processing schemas for detecting the objects and events of the scene.

Bertini et al. [11, 16] have presented an ontology-based framework for semantic video annotation by learning spatio-temporal rules; in their approach, an adaptation of the First Order Inductive Learner to the Semantic Web technologies (FOILS) is used to learn SWRL rule patterns (e.g. Fig. 7) that have been then validated on a few TRECVid 2005 and CAVIAR video events (e.g. Fig. 6). Finally, authors have also contributed to event sharing repositories based on ontologies, with the aim of establishing open platforms for collecting annotating, retrieving and sharing surveillance videos [98, 100] (Fig. 7).

7 Conclusions

The problem of event detection and recognition in videos is acquiring an increasing importance, due to its applicability to a large number of applications, especially considering the problem of human action recognition in video surveillance. Similarly to object recognition there is need to cope with the problem of high variability in lighting variations, geometrical transformation, clutter and occlusion. Moreover, because of the very nature of the problem, it is necessary to consider the temporal dimension of video, requiring thus appropriate features and classification methods to deal with it, and with the variability in the execution of events and actions.

The works presented in this survey have proposed approaches for robust detection and representation of spatio-temporal interest points and motion features, modelling of events and approaches to represent domain knowledge and contextual information of activities and actions. These methods have been applied to several different domains, from sport to surveillance videos, showing promising results. The advances made so far need to be consolidated, in terms of their robustness to real-world conditions and, especially for surveillance applications, there is need of reaching real-time performance.

Acknowledgement This work is partially supported by the EU IST IM3I Project (Contract FP7-222267).

References

1. Akdemir U, Turaga P, Chellappa R (2008) An ontology based approach for activity recognition from video. In: Proc. of ACM multimedia (MM)

2. Arndt R, Troncy R, Staab S, Hardman L, Vacura M (2007) Comm: designing a well-founded multimedia ontology for the web. In: Proc. of int'l semantic web conference
3. Artikis A, Sergot M, Paliouras G (2010) A logic programming approach to activity recognition. In: Proc. of ACM int'l workshop on events in multimedia
4. Assfalg J, Bertini M, Del Bimbo A, Nunziati W, Pala P (2002) Soccer highlights detection and recognition using HMMs. In: Proc. of int'l conference on multimedia & expo (ICME)
5. Assfalg J, Bertini M, Colombo C, Del Bimbo A, Nunziati W (2003) Semantic annotation of soccer videos: automatic highlights identification. Comput Vis Image Underst 92(2–3):285–305
6. Bai L, Lao S, Jones G, Smeaton AF (2007) Video semantic content analysis based on ontology. In: Proc. of int'l machine vision and image processing conference
7. Bai L, Lao S, Zhang W, Jones G, Smeaton A (2007) A semantic event detection approach for soccer video based on perception concepts and finite state machines. In: Proc. intl'l workshop on image analysis for multimedia interactive services (WIAMIS)
8. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2009) Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In: Proc. of int'l conference on image processing (ICIP). Cairo, Egypt
9. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. Multimed Tools Appl 48(2):313–337
10. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Video event classification using string kernels. Multimed Tools Appl 48(1):69–87
11. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Video annotation and retrieval using ontologies and rule learning. IEEE Multimed doi:10.1109/MMUL.2004.4
12. Basharat A, Zhai Y, Shah M (2008) Content based video matching using spatiotemporal volumes. Comput Vis Image Underst 110(3):360–377
13. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) SURF: speeded up robust features. Comput Vis Image Underst 110(3):346–359
14. Bertini M, Del Bimbo A, Nunziati W (2005) Common visual cues for sports highlights modeling. Multimed Tools Appl 27(2):215–218
15. Bertini M, Del Bimbo A, Torniai C, Cucchiara R, Grana C (2007) Dynamic pictorial ontologies for video digital libraries annotation. In: Proc. of ACM int'l workshop on many faces of multimedia semantics (MS)
16. Bertini M, Del Bimbo A, Serra G (2008) Learning ontology rules for semantic video annotation. In: Proc. of ACM int'l workshop on many faces of multimedia semantics (MS)
17. Bloehdorn S, Petridis K, Saathoff C, Simou N, Tzouvaras V, Avrithis Y, Handschuh S, Kompatsiaris I, Staab S, Strintzis M (2005) Semantic annotation of images and videos for multimedia analysis. In: Proc. of European semantic web conference
18. Brand M, Kettnaker V (2000) Discovery and segmentation of activities in video. IEEE Trans Pattern Anal Mach Intell 22(8):844–851
19. Brezeale D, Cook D (2008) Automatic video classification: a survey of the literature. IEEE Trans Syst Man Cybern 38(3):416–430
20. Chao C, Shih HC, Huang CL (2005) Semantics-based highlight extraction of soccer program using DBN. In: Proc. of int'l conference on acoustics, speech, and signal processing (ICASSP)
21. Chen D, Yang J, Wactlar HD (2004) Towards automatic analysis of social interaction patterns in a nursing home environment from video. In: Proc. of int'l workshop on multimedia information retrieval (MIR)
22. Chen M, Hauptmann A, Li H (2009) Informedia @ TRECVID2009: analyzing video motions. In: Proc. of the TRECVID workshop
23. Dasiopoulou S, Mezaris V, Kompatsiaris I, Papastathis VK, Strintzis MG (2005) Knowledge-assisted semantic video object detection. IEEE Trans Circuits Syst Video Technol 15(10):1210–1224
24. Dasiopoulou S, Saathoff C, Mylonas P, Avrithis Y, Kompatsiaris Y, Staab S, Strintzis M (2008) Semantic multimedia and ontologies theory and applications, chapter introducing context and reasoning in visual content analysis: an ontology-based framework. Springer, pp 99–122
25. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: Proc. of int'l workshop on visual surveillance and performance evaluation of tracking and surveillance (VS-PETS)
26. Dousson C, Le Maigat P (2007) Chronicle recognition improvement using temporal focusing and hierarchization. In: Proc. of int'l joint conference on artificial intelligence
27. Dublin Core Metadata Initiative. http://dublincore.org/. Accessed 11 October 2010

28. Ebadollahi S, Xie L, Chang SF, Smith J (2006) Visual event detection using multi-dimensional concept dynamics. In: Proc. of int'l conference on multimedia & expo (ICME)
29. Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
30. Fellbaum C (1998) WordNet: an electronic lexical database, chap 3. A semantic network of English verbs. MIT, Cambridge
31. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
32. Fihl P, Holte M, Moeslund T (2007) Motion primitives for action recognition. In: Proc. of int'l workshop on gesture in human-computer interaction and simulation
33. Francois A, Nevatia R, Hobbs J, Bolles R, Smith J (2005) VERL: an ontology framework for representing and annotating video events. IEEE Multimed 12(4):76–86
34. Garcia R, Celma O (2005) Semantic integration and retrieval of multimedia metadata. In: Proc. of the knowledge markup and semantic annotation workshop
35. Georis B, Mazière M, Brémond F, Thonnat M (2004) A video interpretation platform applied to bank agency monitoring. In: Proc. of intelligent distributed surveillance systems workshop
36. Gruber T (1995) Principles for the design of ontologies used for knowledge sharing. Int J Human-comput Stud 43(5–6):907–928
37. Hakeem A, Shah M (2004) Ontology and taxonomy collaborated framework for meeting classification. In: Proc. of int'l conference on pattern recognition (ICPR)
38. Harte N, Lennon D, Kokaram A (2009) On parsing visual sequences with the hidden Markov model. EURASIP JIVP 2009:1–13
39. Haubold A, Naphade M (2007) Classification of video events using 4-dimensional time-compressed motion features. In: Proc. of ACM international conference on image and video retrieval (CIVR), pp 178–185
40. Hollink L, Little S, Hunter J (2005) Evaluating the application of semantic inferencing rules to image annotation. In: Proc. of int'l conference on knowledge capture
41. Jhuang H, Garrote E, Yu X, Khilnani V, Poggio T, Steele A, Serre T (2010) Automated home-cage behavioral phenotyping of mice. Nature communications doi:10.1038/ncomms.1064
42. Kadir T, Brady M (2001) Saliency, scale and image description. Int J Comput Vis 45(2):83–105
43. Kale A, Sundaresan A, Rajagopalan AN, Cuntoor NP, Roy-Chowdhury AK, Kruger V, Chellappa R (2004) Identification of humans using gait. IEEE Trans Knowl Data Eng 13(9):1163–1173
44. Kennedy L (2006) Revision of LSCOM event/activity annotations, DTO challenge workshop on large scale concept ontology for multimedia. Advent technical report #221-2006-7, Columbia University
45. Kienzle W, Scholkopf B, Wichmann F, Franz MO (2007) How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In: Proc. of 29th annual symposium of the german association for pattern recognition. Springer
46. Kläser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-Gradients. In: Proc. of British machine vision conference (BMVC)
47. Ko T (2008) A survey on behavior analysis in video surveillance for homeland security applications. In: 37th IEEE applied imagery pattern recognition workshop, pp 1–8
48. Kompatsiaris Y, Hobson P (2008) Semantic multimedia and ontologies: theory and applications. Springer
49. Kowalski R, Sergot M (1986) A logic-based calculus of events. New Gener Comput 4(1):67–95
50. Kuettel D, Breitenstein MD, Van Gool L, Ferrari V (2010) What's going on? discovering spatio-temporal dependencies in dynamic scenes. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
51. Laptev I, Lindeberg T (2003) Space-time interest points. In: Proc. of int'l conference on computer vision (ICCV)
52. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
53. Laptev I, Perez P (2007) Retrieving actions in movies. In: Proc. of int'l conference on computer vision (ICCV)
54. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
55. Lavee G, Borzin A, Rivlin E, Rudzsky M (2007) Building Petri nets from video event ontologies. In: Proc. of international symposium on visual computing (ISVC). LNCS, vol 4841. Springer Verlag, pp 442–451

56. Lavee G, Rivlin E, Rudzsky M (2009) Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. IEEE Trans Syst Man Cybern 39(5):489–504
57. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
58. Leslie L, Chua TS, Ramesh J (2007) Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In: Proc. of ACM multimedia (MM)
59. Liu J, Shah M (2008) Learning human actions via information maximization. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
60. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
61. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
62. Luo M, Ma YF, Zhang HJ (2003) Pyramidwise structuring for soccer highlight extraction. In: Proc. of ICICS-PCM
63. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
64. Maillot N, Thonnat M (2008) Ontology based complex object recognition. Image Vis Comput 26(1):102–113
65. Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
66. Mehran R, Moore B, Shah M (2010) A streakline representation of flow in crowded scenes. In: Proc. of European conference on computer vision (ECCV)
67. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
68. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Van Gool L (2005) A comparison of affine region detectors. Int J Comput Vis 65(1/2):43–72
69. Mikolajczyk K, Uemura H (2008) Action recognition with motion-appearance vocabulary forest. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
70. Miller JA, Baramidze G (2005) Simulation and the semantic web. In: Proc. of the winter simulation conference (WSC)
71. Naphade M, Smith J, Tesic J, Chang SF, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91
72. Neumann B, Moeller R (2006) On scene interpretation with description logics. In: Cognitive vision systems: sampling the spectrum of approaches. Lecture notes in computer science, vol 3948. Springer, pp 247–278
73. Nevatia R, Hobbs J, Bolles B (2004) An ontology for video event representation. In: Proc. of the conference on computer vision and pattern recognition workshop (CVPRW)
74. Niebles J, Fei-Fei L (2007) A hierarchical model of shape and appearance for human action classification. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
75. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
76. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: Proc. of European conference on computer vision (ECCV)
77. Oikonomopoulos A, Patras I, Pantic M (2005) Spatiotemporal salient points for visual recognition of human actions. IEEE Trans Syst Man Cybern 36:719
78. Over P, Awad G, Fiscus J, Michel M, Smeaton AF, Kraaij W (2009) TRECVid 2009–goals, tasks, data, evaluation mechanisms and metrics. In: Proc. of the TRECVID workshop. Gaithersburg, USA
79. Paschke A, Bichler M (2008) Knowledge representation concepts for automated SLA management. Decis Support Syst 46(1):187–205
80. Pattanasri N, Jatowt A, Tanaka K (2006) Enhancing comprehension of events in video through explanation-on-demand hypervideo. In: Advances in multimedia modeling. Lecture notes in computer science, vol 4351. Springer, pp 535–544
81. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
82. Sadlier D, O'Connor N (2005) Event detection in field sports video using audio–visual features and a support vector machine. IEEE Trans Circuits Syst Video Technol 15(10):1225–1233

83. SanMiguel J, Martinez J, Garcia A (2009) An ontology for event detection and its application in surveillance video. In: Proc. of int'l conference on advanced video and signal-based surveillance (AVSS)
84. Savarese S, Winn J, Criminisi A (2006) Discriminative object class models of appearance and shape by correlatons. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
85. Savarese S, Del Pozo A, Niebles JC, Fei-Fei L (2008) Spatial-temporal correlatons for unsupervised action classification. In: Proc. of workshop on motion and video computing
86. Scherp A, Franz T, Saathoff C, Staab S (2009) F–a model of events based on the foundational ontology DOLCE+DnS ultralight. In: Proc. of int'l conference on knowledge capture (K-CAP)
87. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proc. of int'l conference on pattern recognition (ICPR)
88. Scovanner P, Ali S, Shah M (2007) A 3-Dimensional SIFT descriptor and its application to action recognition. In: Proc. of ACM multimedia (MM)
89. Seidenari L, Bertini M (2010) Non-parametric anomaly detection exploiting space-time features. In: Proc. of ACM multimedia (MM)
90. Shet V, Harwood D, Davis L (2005) Vidmap: video monitoring of activity with prolog. In: Proc. of IEEE int'l conference on advanced video and signal-based surveillance (AVSS)
91. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proc. of int'l conference on computer vision (ICCV)
92. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proc. of int'l workshop on multimedia information retrieval (MIR)
93. Snidaro L, Belluz M, Foresti G (2007) Domain knowledge for surveillance applications. In: Proc. of int'l conference on information fusion
94. Snoek C, Worring M (2005) Multimodal video indexing: A review of the state-of-the-art. Multimed Tools Appl 25(1):5–35
95. Tran SD, Davis LS (2008) Event modeling and recognition using Markov logic networks. In: Proc. of European conference on computer vision (ECCV)
96. Tsinaraki C, Polydoros P, Kazasis F, Christodoulakis S (2005) Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content. Multimed Tools Appl 26(3):299–325
97. TV Anytime Forum. http://www.tv-anytime.org/. Accessed 11 October 2010
98. Vezzani R, Cucchiara R (2010) Video surveillance online repository (ViSOR): an integrated framework. Multimed Tools Appl 50(2):359–380. http://www.openvisor.org
99. Viola PA, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
100. Wang Xj, Mamadgi S, Thekdi A, Kelliher A, Sundaram H (2007) Eventory—an event based media repository. In: Proc of the int'l conference on semantic computing (ICSC)
101. Wang F, Jiang YG, Ngo CW (2008) Video event detection using motion relativity and visual relatedness. In: Proc. of ACM multimedia (MM)
102. Willems G, Tuytelaars T, Van Gool L (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proc. of European conference on computer vision (ECCV)
103. Winder SAJ, Hua G, Brown M (2009) Picking the best DAISY. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
104. Wong SF, Cipolla R (2007) Extracting spatiotemporal interest points using global information. In: Proc. of int'l conference on computer vision (ICCV)
105. Wong SF, Kim TK, Cipolla R (2007) Learning motion categories using both semantic and structural information. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
106. 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
107. Xu P, Xie L, Chang SF, Divakaran A, Vetro A, Sun H (2001) Algorithms and system for segmentation and structure analysis in soccer video. In: Proc. of int'l conference on multimedia & expo (ICME)
108. Xu G, Ma YF, Zhang HJ, Yang S (2003) A HMM based semantic analysis framework for sports game event detection. In: Proc. of IEEE int'l conference on image processing (ICIP). Barcelona, Spain
109. Yang J, Hauptmann AG (2006) Exploring temporal consistency for video analysis and retrieval. In: Proc. of int'l workshop on multimedia information retrieval (MIR)
110. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proc. of int'l workshop on multimedia information retrieval (MIR)

111. Zhan B, Monekosso D, Remagnino P, Velastin S, Xu LQ (2008) Crowd analysis: a survey. Mach Vis Appl 19:345–357
112. Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vis 73(2):213–238
113. Zhou X, Zhuang X, Yan S, Chang SF, Hasegawa-Johnson M, Huang T (2008) SIFT-bag kernel for video event analysis. In: Proc. of ACM multimedia (MM), pp 229–238

Lamberto Ballan is a PhD student at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence, Italy. His main research interests focus on multimedia information retrieval, pattern recognition, computer vision and machine learning. He has a laurea degree in computer engineering from the University of Florence. Contact him at ballan@dsi.unifi.it.

Marco Bertini is an assistant professor in the Department of Systems and Informatics at the University of Florence, Italy. His research interests include content-based indexing and retrieval of videos and Semantic Web technologies. Bertini has a PhD in electronic engineering from the University of Florence. Contact him at bertini@dsi.unifi.it.

Alberto Del Bimbo is a full professor of computer engineering at the University of Florence, Italy, where he is also the director of the Master in Multimedia Content Design. His research interests include pattern recognition, multimedia databases, and human-computer interaction. He has a laurea degree in Electronic Engineering from the University of Florence. Contact him at delbimbo@dsi.unifi.it.

Lorenzo Seidenari is a PhD student at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence, Italy. His research interests include spatio-temporal features, human action recognition, anomaly detection and machine learning applied to video understanding. Lorenzo Seidenari has a laurea degree in computer engineering from the University of Florence. Contact him at seidenari@dsi.unifi.it.

Giuseppe Serra is a postdoc at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence, Italy. His research interests include multiple-view geometry, self-calibration and 3D reconstruction, and video understanding based on statistical pattern recognition and ontologies. Serra has a laurea degree in computer engineering from the University of Florence. Contact him at serra@dsi.unifi.it.