Multimed Tools Appl (2016) 75:297–331
DOI 10.1007/s11042-014-2292-8

Towards large-scale multimedia retrieval enriched by knowledge about human interpretation

Retrospective survey

Kimiaki Shirahama · Marcin Grzegorzek

Received: 11 March 2014 / Revised: 22 August 2014 / Accepted: 19 September 2014 /
Published online: 5 October 2014

Abstract Recent Large-Scale Multimedia Retrieval (LSMR) methods seem to heavily rely on analysing a large amount of data using high-performance machines. This paper aims to warn this research trend. We advocate that the above methods are useful only for recognising certain primitive meanings, knowledge about human interpretation is necessary to derive high-level meanings from primitive ones. We emphasise this by conducting a retrospective survey on *machine-based* methods which build classifiers based on features, and *human-based* methods which exploit user annotation and interaction. Our survey reveals that due to prioritising the generality and scalability for large-scale data, knowledge about human interpretation is left out by recent methods, while it was fully used in classical methods. Thus, we defend the importance of *human-machine cooperation* which incorporates the above knowledge into LSMR. In particular, we define its three future directions (cognition-based, ontology-based and adaptive learning) depending on types of knowledge, and suggest to explore each direction by considering its relation to the others.

Keywords Large-scale multimedia retrieval · Human-machine cooperation · Machine-based methods · Human-based methods

1 Introduction

Confucius who is an ancient Chinese social philosopher, said "reviewing what you have learned and learning anew, you are fit to be a teacher". This means an approach to discover new things based on the study of the past. In this spirit, we conduct a retrospective

K. Shirahama (✉) · M. Grzegorzek
Pattern Recognition Group, University of Siegen, Hoelderlinstrasse 3, 57076 Siegen, Germany
e-mail: kimiaki.shirahama@uni-siegen.de

M. Grzegorzek
e-mail: marcin.grzegorzek@uni-siegen.de

survey on *Large-Scale Multimedia Retrieval* (LSMR) which has been receiving much research attention for more than twenty years. LSMR is the technique for analysing a large amount of multimedia data to efficiently find interesting and relevant ones. In other words, LSMR can be regarded as a classification problem to discriminate between relevant and irrelevant data to a query. As described in many literature [28, 113, 117, 121], the most challenging issue is the *semantic gap* which is the lack of coincidence between automatically extractable features (e.g., colour, edge and motion) and human-perceivable semantic meanings.

First of all, by referring to Fig. 1, let us define meanings that LSMR needs to identify. Since events are widely-accepted access units to multimedia data, we decompose semantic meanings based on basic aspects of event descriptions [101, 144]. As shown in Fig. 1a, we organise meanings using three components, *concept*, *event* and *context*. By applying these to [101, 144], concepts form the participation (or informational) aspect of objects in an event. That is, the event is derived by relating multiple concepts. Contexts are the collection of part-of, causal and correlation aspects among events.

More formally, we define concepts as textual descriptions of meanings that can be perceived from images, shots or videos, such as objects like *Person* and *Car*, actions like *Walking* and *Airplane_Flying*, and scenes like *Outdoor* and *Nighttime*. In other words, concepts are the most primitive meanings for multimedia data, and used in many state-of-the-art retrieval systems [80, 114]. Below, concept names are written in italics to distinguish them from the other terms. An event is a higher-level meaning derived from the interaction of objects at a specific situation [50, 111]. In our case, it is defined by the combination of concepts. For example, in Fig. 1b, *Shot 1* shows *Cheese*, *Meat*, *Sausage* and *Grill*, from which the event "burning the first three things" is derived. *Shot 2* displays *Hand*, *Food_Turner*, *Bread*, *Cheese* and so on, where the event "putting *Cheese* etc. on *Bread*" is formed based on movements of these concepts. Furthermore, as depicted by the bold line arrow in Fig. 1a,

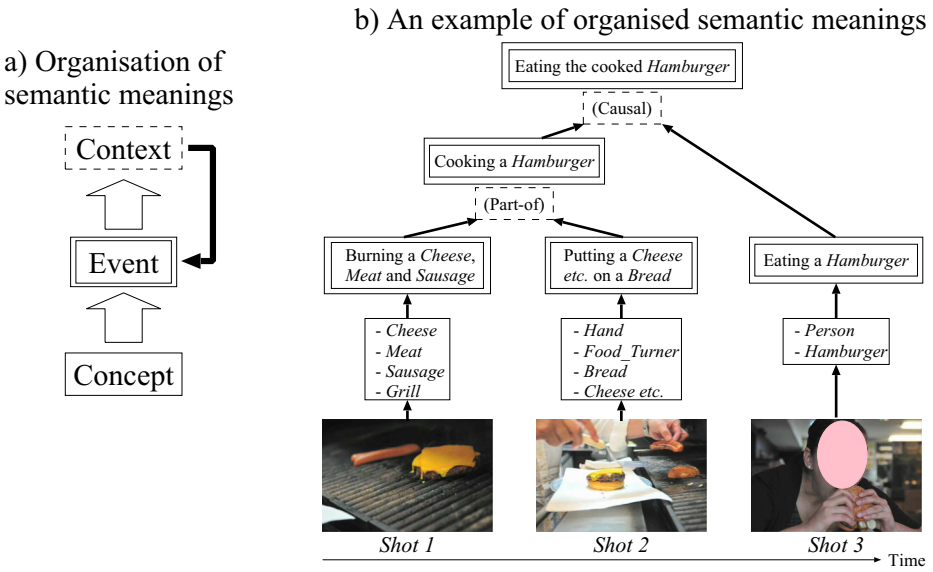

Fig. 1 An illustration of decomposing meanings based on concepts, events and contexts

contexts are used to recursively define higher-level events based on part-of, causal and correlation relations among lower-level ones.[1] In Fig. 1b, based on the part-of relation, events in *Shot 1* and *2* are combined into the higher-level event "cooking a *Hamburger*". This event and the one in *Shot 3* ("eating a *Hamburger*") are further abstracted into "eating the cooked *Hamburger*". Also, the correlation relation is used to connect two 'weakly-related' events, such as those which occur in separate locations but at the same time [101]. We consider the above organisation of meanings based on concepts, events and contexts as the final goal of LSMR.

To make the following discussions simple and clear, we adopt two policies: First, we use an *example* to indicate a single unit of multimedia data, such as image, shot, video and audio. When the discrimination among these data formats is not important, we use examples as their abstract name. Second, by drawing an analogy with Content-Based Image Retrieval (CBIR) in [24], we define LSMR as any technology that, in principle, helps to organise a large-scale multimedia data. Hence, LSMR in this paper includes technologies such as object detection/recognition, image/video/audio classification, browsing, summarisation and so on.

Our motivation for this survey paper is attributed to the fact that, the current LSMR owes much to the availability of large-scale data and the enhancement of machine performance. The underlying framework remains the classical *machine learning* approach. Roughly speaking, a classifier is built by analysing *training examples* each annotated with the presence or absence of a certain meaning. The former training examples, *positive examples*, serve as representatives of examples relevant to the meaning, while the latter ones, *negative examples*, represent irrelevant examples. Here, examples have significantly different visual appearances (features) depending on various changing factors (e.g., camera techniques and shooting environments). Thus, by analysing a larger number of training examples with a high-performance machine, the classifier can accurately distinguish *test examples* relevant to the meaning from the others irrespective of changing factors.

However, the success of the above machine learning approach does not mean that, machines which are fed with a large number of training examples, learn a classifier the way humans do. This approach is only useful for certain types of concepts (see Section 3.2). Furthermore, an event is derived from the combination of concepts, and what is more, a context is composed of multiple events. Hence, examples relevant to the event or context incur much more diverse visual appearances than examples relevant to a concept. Thus, the detection of the former requires much more training examples than that of the latter. But, since events and contexts are very specific meanings, collecting many training examples is impractical. Thus, knowledge about human interpretation of semantic meanings is necessary to derive events from automatically detectable concepts, and further deduce contexts from events. In other words, this knowledge is used to effectively cover the huge diversity of visual appearances, connected to examples relevant to an event or context. Therefore, this paper defends the importance of *human-machine cooperation* approaches which enrich LSMR by knowledge about human interpretation.

[1] In this paper, contexts only include relations which are obtained from multimedia data themselves, and exclude external data like geo-tags and Web documents.

2 Overview of our retrospective survey

Our survey approach is significantly different from those of existing papers in fields of multimedia retrieval and understanding. Most survey papers adopt a *progressive* approach to derive future research directions from the progress of component technologies. Recent papers [15, 24, 50, 62, 68, 114] mainly reviewed the following four component technologies, (1) feature extraction, representation and transformation methods, (2) retrieval methods based on knowledge bases, machine learning techniques, similarities in terms of features, and data mining methods, (3) user interaction methods such as query specification, browsing (visualisation) and feedback, and (4) benchmark datasets for performance evaluation. Then, the above papers suggest future problems that should be further explored or should receive more attention, such as improvement of component technologies, design of application-oriented (human-centric) interfaces, scalability with both high-performance computing and algorithm sophistication, synergy between different modalities like text, image, video and audio, and utilisation of user-generated Web data like tagged images and videos.

Compared to the existing papers, we conduct a survey in a *retrospective* approach. By tracing the progress of LSMR, we detect missing links from recent approaches, which were addressed by classical approaches. Roughly speaking, due to the large data size that is unmanageable by humans, researchers tend to leave LSMR just to machines, where no knowledge about human interpretation is used. Thus, we argue the importance of human-machine cooperation which explicitly incorporates knowledge into machine-based approaches. In other words, researchers have already established large-scale labelled examples from which we now need to extract knowledge and utilise it, in order to achieve retrieval for high-level meanings. This means the 'return' to classical approaches, but we also need to consider the much larger and more structured knowledge in the future LSMR.

As depicted by the time axis in Fig. 2, we review existing LSMR methods in chronological order. We classify them into three categories, *machine-based*, *human-based* and human-machine cooperation. Machine-based LSMR does not explicitly utilise knowledge about human interpretation. This category began with *syntactic* methods using pre-defined templates of features based on specific camera and editing techniques in a particular video

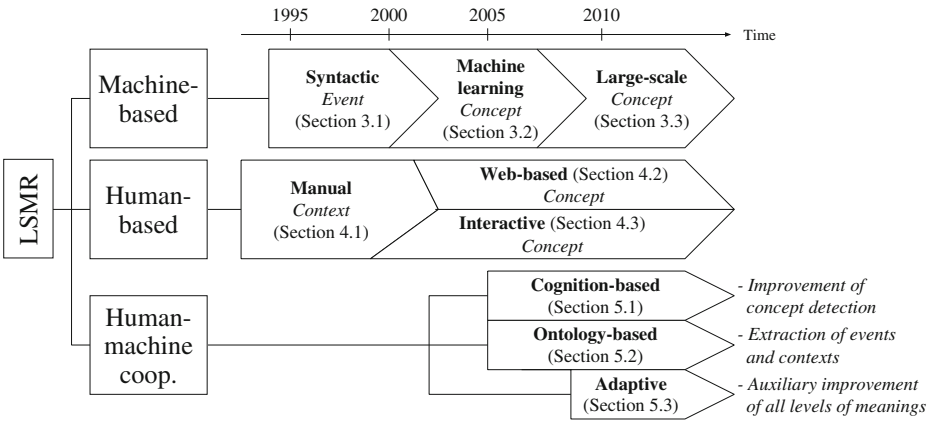

Fig. 2 Overview of our retrospective review of existing LSMR methods

domain. In Fig. 2, we locate their starting point before 1995 by regarding one of the earliest method by Zhang et al. in 1993 [162]. In Fig. 2, the italic term (i.e., *Concept*, *Event* or *Context*) under a method name represents the level of meaning that can be detected. We assume that according to the compositional relation in Fig. 1a, methods which can detect events (or contexts) have the capability to expose concepts (or concepts and events). Syntactic methods support retrieval of events over shot sequences, such as goal events in football videos and conversation events in movies. However, syntactic methods lack the generality because pre-defined templates are sensitive to changes in shooting environments and shot concatenations. Thus, since around 2000, the research focus has shifted to *machine learning* methods which statistically build a classifier using training examples. As discussed before, these methods are only effective for concepts. As the effectiveness of using a large number of training examples together with high-dimensional features [23, 79] became well-known, after 2005, researchers started to develop *large-scale* methods which improve the scalability of machine learning ones.

Human-based LSMR is supported by a human, but machine and human are independent of each other. The earliest *manual* methods perform retrieval based on manual annotation. In Fig. 2, their starting point is set before syntactic ones, because multimedia retrieval was originally explored as a database problem (e.g., [146]) where examples are manually indexed to flexibly respond to queries of all levels of meanings. However, considering the labour for manual annotation, from the early 2000s, *Web-based* methods have received attention where users on the Web collaboratively annotate large-scale data [65]. Note that annotation in such methods has to be done easily so that many users can participate in it. In consequence, Web-based methods only support simple concept-level retrieval. Meanwhile, human-based LSMR includes *interactive* methods in which a human provides additional training examples based on the current retrieval result, as represented by relevance feedback developed in the late 1990s [98]. Considering the practical use, feedbacks should be done efficiently. Since it takes time to check shots or videos, interactive methods are suitable for concepts where their presence or absence can be quickly judged from single images. It should be noted that interactive methods do not affect retrieval algorithms, but just provide additional data to tune parameters. Compared to this, human-machine cooperation addresses the collaboration of humans and machines at the algorithm level.

As can be seen from Fig. 2, although early methods supported retrieval for high-level meanings (events and contexts), recent ones can only perform concept-level retrieval because of preferring to the generality and scalability for large-scale data, and the usability for users. Concepts are not so useful for practical applications because they are too primitive (or general) to identify examples that users want to retrieve. Therefore, the future LSMR should address the extraction of events and contexts while improving the concept detection performance. Human-machine cooperation is promising to achieve these goals and should deserve more attention. As shown in Fig. 2, we describe three types of methods. First, *cognition-based* methods utilise knowledge about human visual system. These aim to improve concept detection by implementing the mechanism of how human brains process visual information. Second, *ontology-based* methods make use of knowledge about human inference for high-level meanings, in order to derive events and contexts from concepts. In Fig. 2, we locate the starting points of the above methods from 2005, when the importance of a standardised set of concepts became well-known [80]. Last, *adaptive leaning* methods utilise knowledge about human learning, and adaptively control components of various retrieval algorithms. Thus, their role is the auxiliary improvement of detecting all levels of meanings. We set the starting point of adaptive learning to 2009, because, to the best of our knowledge, one of the earliest methods was developed in [12]. We hope that the discussion

about the human-machine cooperation reminds researchers of the importance of knowledge for interpreting semantic meanings.

3 Machine-based LSMR

In this section, we first describe classical syntactic methods and their disadvantages. Then, we review machine learning methods which overcome the disadvantages of syntactic ones. Finally, recent works are presented where the scalability of machine learning approaches is improved.

3.1 Syntactic approaches

The easiest approach is to utilise prior knowledge about the structure (syntax) of videos. This is only possible when the retrieval is limited in a specific genre of videos. For example, in news videos, an event[2] that represents one news topic starts with a shot where an anchor person appears, and ends with another shot where the anchor person appears again [161, 162]. In addition, an interview event is characterised by a sequence of shots, where shots showing an interviewer alternate with shots showing an interviewee [88]. To detect these events, researchers first construct a graph where each node represents a group of visually similar shots, and each edge represents the transition between two groups of shots [88, 161]. Events are then detected by extracting cycles which are connected to the node of shots showing the anchor person.

In a sports video, an event corresponding to one move in a game starts with a specific shot [164]. For example, in a baseball video, an event starts with a shot taken behind the pitcher. And, when the batter hits out the ball, a camera follows the flight of the ball in the next shot. In an American football video, each play starts with the formation where players line up on two sides of the ball. Also, goal events in ball game videos are characterised by a score change on the score caption, followed by audience's cheering and applause [167]. Based on the above heuristics, each event is modelled using a pattern which represents a sequence of characteristic features [167], or a Hidden Markov Model (HMM) [4]. Then, the event in an unknown video is detected by finding sequences of shots, which match the pattern or HMM.

Movie directors and editors use *film grammar* which consists of practical rules to concentrate viewer's attention on the story of a video [77]. For example, thrilling events are presented with a fast transition of shots with very short durations in order to emphasise the thrilling mood. On the other hand, romantic events are created by concatenating shots with very long durations, where person's emotions and actions are thoroughly presented. In addition, a conversation between two persons is displayed by alternating shots showing the two persons one after another. Based on film grammar, conversation, suspense and action events are extracted using sequential patterns [154] or Finite State Machines (FSMs) [160]. Such a pattern or FSM represents a characteristic sequence of features, such as shot duration, motion, audio, and repetition of visually similar shots. Also, in [1], a tempo at which a

[2]Depending on literature, a sequence of shots that are coherent to a certain location, action or theme, is named as a different term like scene [4, 78, 160], event [104, 106, 167], or story section [1]. In this paper, such a sequence is called an event based on Fig. 1.

viewer perceives meanings (e.g., haste and calm) is computationally defined based on shot durations and camera movements. Dramatic events are then detected by extracting gradual and sharp changes of the tempo. Furthermore, violent events are detected based on multiple heuristically defined features, such as motion, shot duration, flame-colour, blood-colour and sudden increase in audio energy [78].

In a surveillance video recorded by a fixed camera, the background frame can be defined as the one where no object appears. Based on this, the movement of an object can be easily captured by computing the difference between a video frame and the background frame. Thus, an event where an object actively moves is detected as a sequence of video frames which have large differences to the background frame [86]. Also, it is assumed that most events taken by a surveillance camera are normal and anomalous ones are very rare. Hence, by grouping video frames into clusters of similar frames, clusters including a large number of frames and clusters including a small number of frames represent normal and anomalous events, respectively [166].

The above syntactic methods can only process a limited number of a-priori known queries. However, users issue a variety of queries which cannot be assumed in advance. To overcome this, some research effort has been made on *video data mining* where videos are analysed using data mining techniques that extract previously unknown, interesting patterns in underlying data [104, 106]. As a result, patterns for retrieving a variety of events can be extracted. Specifically, we developed a method which extracts sequential patterns for associating adjacent shots related to a certain event [106]. Such sequential patterns are extracted by connecting statistically correlated features in adjacent shots. In addition, we also devised a method which extracts patterns for characterising 'topics' [104]. A topic is an event showing an interesting action like fighting, chasing or kissing. It is assumed that topics are not presented by normal editing patterns but by abnormal patterns, because the latter ones have much more impact on viewers than the former ones. In [104], a probabilistic time series segmentation is developed to extract abnormal patterns, each of these showing a certain person appearing in continuous shots with abnormally long or short durations.

However, even using video data mining methods, it is practically impossible to prepare all patterns which can respond to a variety of queries. In addition, pre-specified patterns or retrieval models lack the generality because it is difficult to assume a diversity of camera and editing techniques, that can be used to present an event. Thus, the research focus was shifted to a more general and flexible approach, where a user represents a query by providing training examples, based on which a retrieval model is constructed on the fly. The next section provides this kind of machine learning methods.

3.2 Machine learning approaches

Machine learning is applied to LSMR as *Query By Example* (QBE) where a user first provides some training examples as a query, then a classifier is constructed using them [45, 94]. The classifier distinguishes test examples relevant to the query from the others. Classical QBE methods search for test examples that are the most similar to given positive examples. The following two research topics have been assiduously explored. The first is the development of good similarity measures between positive and test examples. Many similarity measures such as histogram-based measure [46], psychology-based measure [67], a measure based on weighted graph matching [92], a measure based on longest common subsequence (LCS) [54], were developed. The other topic is the speed-up of the similarity calculation. For example, Kashino et al. developed the method that avoids unnecessary similarity calculation by estimating the upper bound of similarity [53], and Yuan et al. devised

the two-phase hierarchical method that first computes a coarse similarity on sub-sampled video frames, and then verifies the similarity using fine audio features [156].

One big disadvantage of classical QBE methods is that they use *global features* which represent overall characteristics of an example, such as a colour histogram indicating overall colours, and a texture vector expressing overall responses to different filters. Such global features cannot capture the detailed content of the example. For this, in the late 1990s, Schmid and Mohr proposed to represent the example as a collection of *local descriptors*, each of which represents the characteristic of a local region [102]. To avoid confusion, we define a descriptor as the representation of a local region [163], and a feature as the representation of an example based on a set of local descriptors. Figure 3 illustrates feature extraction using local descriptors. First, as depicted by yellow circles in Fig. 3a, local descriptors are sampled from small regions in an example. Then, the example is represented by a feature which represents the distribution of sampled local descriptors. The feature (i.e., distribution) in Fig. 3b indicates that the example contains many descriptors similar to the one marked with (1), and few descriptors similar to (2). Such a feature reflects the detailed content of the example. In particular, by sampling a large number of local descriptors, the feature become robust to shape deformations and occlusions. For example, even if the car in Fig. 3a is partially masked by other objects, local descriptors that characterise its visible parts like a wheel, window or headlight are included in the feature. Below, we review existing QBE methods in terms of local descriptors, features and classifiers.

Many local descriptors have been proposed to capture different characteristics of a local region. The most popular one is a *Scale-Invariant Feature Transform* (SIFT) descriptor which represents the shape in a local region, reasonably irrespective of changes in illumination, rotation, scaling and viewpoint [69]. Sande et al. developed SIFT descriptors that are defined in different colour spaces and have unique invariance properties for lighting conditions [132]. Furthermore, local descriptors are defined around trajectories, each of which is obtained by tracking a point in a video [141]. The resulting local descriptors represent the displacement of a tracked point, the derivative of that displacement, and edges around a trajectory. Also, *Speeded-UP Robust Features* (SURF) descriptors are similar to SIFT descriptors, but can be efficiently computed based on the integral image structure which quickly identifies the sum of pixel values in any image region [9].

The main research theme for feature extraction is to accurately represent the distribution of local descriptors sampled from an example. The simplest approach, called *Bag of Visual Words* (BoVW), represents the distribution as a collection of characteristic local descriptors, namely *visual words* [23]. A set of local descriptors are firstly grouped into clusters where each cluster centre is a visual word. Then, each local descriptor extracted from an example is assigned to the most similar visual word. As a result, the example is

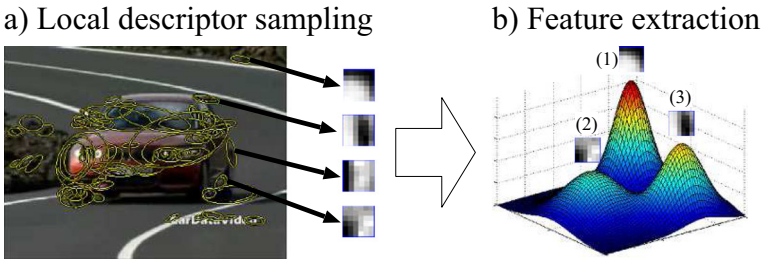

Fig. 3 An illustration of feature extraction based on local descriptors

represented as a histogram which represents the frequency of each visual word. Many extensions of BoVW have been proposed, such as soft assignment which extracts a smoothed histogram by assigning each local descriptor to multiple visual words based on kernel density estimation [132], sparse coding which represents the distribution of a large number of base functions used to sparsely approximate local descriptors [151, 155], Gaussian Mixture Model (GMM) supervector which estimates the distribution of local descriptors using a GMM [44], Fisher vector encoding which considers the first and second order differences between the distribution of local descriptors and the reference distribution [93], and Vector of Locally Aggregated Descriptors (VLAD) which concatenates vectors each representing the accumulated difference of a visual word to the assigned local descriptors [5, 47].

Since a feature which precisely represents the distribution of local descriptors is necessarily high-dimensional, a classifier effective for high dimensional data is used in QBE. Typically, a *Support Vector Machine* (SVM) is used [23, 49, 107, 132, 163] because its 'margin maximisation' principle can extract a well-generalised classification boundary between positive and negative examples in the high-dimensional feature space [134]. It should be noted that only positive examples are provided in QBE. Regarding this, Natsev et al. proposed to use randomly sampled examples as negative by assuming that only a small number of examples in the database are relevant to a query [81]. That is, almost all of the randomly selected examples are irrelevant and serve as negative. This approach works reasonably well and has been used in many existing works [83, 107, 116]. In addition, to cover a diversity of examples relevant to a query, bagging and random subspace are used to combine multiple SVMs which are built using subsets of randomly selected training examples, and subsets of randomly selected feature dimensions, respectively [81, 125, 150]. Such SVMs characterise different portions of relevant examples. In this context, we proposed a method using *rough set theory* which is a set-theoretic classification approach for extracting rough descriptions of a class from imprecise (or noisy) data [107]. Specifically, our method extracts classification rules each of which represents an SVM combination to correctly identify a different subset of positive examples. By accumulating relevant examples with such classification rules, a variety of relevant examples can be accurately covered.

Compared to classical syntactic methods, the above machine learning methods have much more generality because examples relevant to a query can be retrieved with significantly higher accuracy, regardless of video genres, camera techniques and shooting environments. However, machine learning methods are useful only for concepts corresponding to 'basic categories' of meanings, such as *Person*, *Car* and *Building*. Even though visual appearances of each basic category significantly vary, these are apparently different from those of the other basic categories [26, 60]. Compared to this, for concepts corresponding to 'subordinate categories' like *Rider*, *Driver* and *Factory_Worker* of *Person*, and *Bus*, *Van* and *Truck* of *Car*, their visual appearances can be distinguished only by small localised regions, or considering the relation to surrounding concepts. In addition, since an event (or context) involves multiple concepts, examples relevant to it incur a much larger variance of features than the one of examples relevant to a concept. Therefore, knowledge about human interpretation is necessary to appropriately detect concepts for subordinate categories, events and contexts.

3.3 Using large-scale data

This section presents methods for scaling up machine learning methods to large-scale data. We mainly focus on concept detection where a large number of training examples are available. Recently, there are several worldwide competitions, such as TRECVID [111]

and PASCAL VOC [90], where concept detection methods developed in different research institutes are compared using large-scale benchmark data. These competitions have been promoting the improvement of concept detection methods.

One of the most important issues in concept detection is that different camera techniques and shooting environments cause visually diverse examples where a certain concept appears. To cover such a diversity, a large number of training examples are required. In general, the detection performance is proportional to the logarithm of the number of positive examples, although each concept has its own complexity of detection [79]. This means that 10 times more positive examples improve the performance by 10 %. In an extreme case, 80 million training images yield accurate detection performance [129]. Furthermore, using two billion images, specific concepts such as celebrities, consuming electronics and landmarks can be detected accurately [140]. Based on the importance of the number of training examples, researchers have developed online systems where many users on the Web collaboratively annotate a large number of examples as positive or negative [6, 135].

Another important issue is sampling of local descriptors. Algorithms for extracting local descriptors generally consist of two modules, *region detector* and *region descriptor* [163]. The former detects regions useful for characterising concepts, and the latter represents each of the detected regions as a vector. A concept is shown in significantly different regions, and in videos, it does not necessarily appear in all video frames. Considering such 'unclear' concept appearances, it is effective to exhaustively sample local descriptors in both the spatial and temporal dimensions. Indeed, the performance is improved as the number of sampled local descriptors increases [84]. Moreover, Snoek et al. compared two methods. One extracts features only from one video frame in each shot (one shot contains more than 60 frames), and the other extracts features every 15 frames [115]. They found out that the latter exceeds the former by 7.5 to 38.8 %.

Although a large number of training examples and exhaustively sampled local features are immensely important for accurate concept detection, processing them requires high computational costs. Many methods for reducing these costs have been developed. They can be classified into two types, 'high-performance computing' and 'algorithm sophistication'. The first type parallelises the classifier training/testing process and the feature extraction using special hardware, such as a cluster consisting of multiple PCs [2, 150], multicore CPU [22] and General-Purpose computing on Graphics Processing Units (GPGPU) [18, 133]. The second type includes a fast SVM training method that iteratively solves sub-problems consisting of the most problematic training examples [29], a fast SVM training method that iteratively solves simple one-variable sub-problems [42], a fast SVM training and test method that efficiently computes similarities (kernel values) by sorting dimension values of each example [73], a fast SVM test method by hashing SVM parameters and the feature of each test example [66], and a fast feature extraction method by organizing the distribution of local features into a tree structure [44].

We also developed a fast SVM training/test method and a fast feature extraction method based on matrix operation [105]. The former re-formulates similarity computation, which enables batch computation of similarities among many examples. The latter re-formulates probability density computation, so that probability densities of many local descriptors can be computed in a batch. Based on these, SVM training/test and feature extraction become about 10–37 and 5–7 times faster than the normal implementation, respectively. By processing a large number of training examples and exhaustively sampled local descriptors using the above methods, we achieved the highest performance in TRECVID 2012 Semantic Indexing (light) task, which is one of the most predominant worldwide competitions on video analysis and retrieval [105].

Finally, as described above, much research effort has been invested to scale up machine learning methods to large-scale data, still the underlying framework remains the same. In other words, researchers prioritise the generality and scalability of a method, so that the same method can be used to search large-scale data in terms of a variety of queries. Of course, improving the generality and scalability is very important. But, we think that, the intensive favour to it is one reason why the mechanism of recent LSMR methods has become completely different from the mechanism of human's semantic meaning interpretation.

4 Human-based LSMR

This section first presents classical human-based LSMR methods based on manual annotation. Recent methods are then described where manual annotation is conducted collaboratively by users on the Web. Finally, we review interactive methods that enable users to interactively refine retrieval results.

4.1 Manual annotation

In classical manual video annotation methods, videos are manually annotated with text descriptions. The following three issues are mainly addressed [123]:

1. Identification of meaningful segments: Videos are known as *continuous media* where sequences of media quanta (i.e., video frames and audio samples) convey semantic meanings when continuously played over time [36]. Hence, any segment of a video can become a meaningful unit.
2. Annotation that should be provided: A video contains too many meanings ranging from low-level ones like colour and shape to high-level ones like event and context. Thus, it is difficult to annotate the video with all the meanings contained in it.
3. Discrepancy between annotation and user expectation: This focuses on segments that are annotated and segments that are expected to be retrieved by users. For example, one intuitive answer to the query "two persons A and B are talking to each other" is a shot annotated with both A's and B's presences. However, a sequence of shots can be another answer where shots annotated only with A's presence and shots annotated only with B's presence are repeated one after the other. Thus, dynamic organisation of annotated shots (segements) is required to correctly respond to queries.

In accordance with these, we present several manual annotation methods.

Weiss proposed the algebraic video data model where text descriptions are organised in a nested hierarchical way by considering their temporal relationships, such as overlapping and inclusion [143]. This facilitates an easy way to attach different meanings to the same segment, and construct compound meanings from annotated meanings using algebraic operations like union, intersection, concatenation and so on. Oomoto and Tanaka developed Object-oriented Video Information Database (OVID) where a segment and text descriptions are regarded as a video object and attribute values, respectively [87]. Such attribute values of a video object are inherited by another object based on their temporal inclusion relationship. This way, text descriptions are shared among video objects, so that the manual annotation effort is significantly reduced.

Uehara et al. proposed an approach which represents the story of a video using a binary tree, called a story graph [130]. In this graph, each node represents the relation (e.g., sequential, physically-causal and psychologically-causal) between two successive segments, and

edges are labelled with semantic constraints. This enables users to retrieve arbitrary-length scenes specified by natural language, and retrieve causes or consequences of queries based on causal relationships. Zettsu et al. developed a time-stamped authoring graph where each node represents a text description at a certain point of time in a video, and two nodes are connected if they have a strong semantic correlation based on co-occurrences of words in text descriptions [158]. Given a query, a segment is retrieved as the minimal subgraph which consists of nodes containing words in the query. This way, meaningful segments are dynamically determined depending on issued queries.

Pattanasri et al. developed a method using a knowledge base (ontology) about contexts [91]. This knowledge base represents relationships among verbs, such as "kill" implies "die". Thereby, video segments that are related in terms of causes and effects of person's actions, can be linked together and retrieved as a whole. François et al. developed an extensible and hierarchical framework for representing events in videos [33]. Here, complex events are constructed from simpler ones by operations, such as sequencing, iteration and alternation, which are defined in a knowledge base. Like this, various complex events can be defined only using relatively few primitive events.

Since the above approaches require expensive manual annotation cost, research focus has been shifted to machine learning methods in Section 3.2 and the ones described in the next section. It should be noted that very flexible context-level retrieval is possible based on laborious manual annotation.

4.2 Web-based annotation

Many Web-based annotation systems have been developed to distribute manual annotation of large-scale multimedia data to many users on the Web. To achieve this, the *usability* and *quality* of annotation should be addressed. The usability means whether users can easily annotate examples or not. If this is insufficient, it cannot be expected that many users participate in annotation. Regarding the quality, meaningless annotation may be provided by malicious users or operation mistakes.

The IBM research group developed a system for annotating a large number of shots with concepts' presences or absences [65, 135]. To improve the usability, users are allowed to customise their annotation styles, such as the number, size, and layout of shots displayed per page, using mouse and/or keyboard, and annotating one or more concepts at a time. In addition, the system informs a user of how difficult the annotation of each concept is based on the disagreement with past annotations by different users, so that the annotation quality improves. In [6], the system in [135] is extended using active learning, in order to preferentially annotate shots that are promising for improving the classifier of a concept (see Section 4.3 for more detail about active learning).

Russell et al. developed LabelMe which is a Web-based system for annotating object (concept) regions in images [99]. Given an image, the user labels an object region by creating a polygonal region by mouse, then types the object name. To improve the usability and maintain the consistent annotation, the researchers considered several extensions, such as the lexical knowledge base (WordNet) for expanding and disambiguating freely typed object names, and the object relation for suggesting candidate objects where their regions frequently overlap a user-specified region.

However, the above systems do not consider the motivation of users. In other words, regular users on the Web are unlikely to volunteer to annotate when no benefit or no reason is given. In consequence, only researchers participate in annotation, which makes it difficult to collect large-scale annotation. Von Ahn and Dabbish proposed a *Games With A Purpose*

(GWAP) approach where users play a game, and as a side effect, a computationally difficult task is solved [136, 137]. More concretely, users play a fun game without knowing that they conduct image annotation. Owing to the motivation that users want to have fun, as of July 2008, 200,000 users contributed to assigning more than 50 million labels to images on the Web [137].

The first game based on the GWAP approach is the ESP game where randomly paired users are first given the same image, then each user guesses a label that another user is likely to provide [136, 137]. If labels provided by both users agree, they get a certain number of points, and the next image is given. Like this, users are encouraged to get more points and play the ESP game many times. Since users know nothing and cannot communicate with each other, the easiest way for them to earn points is to provide labels relevant to given images. Thus, annotation data obtained by the EPS game are likely to be meaningful. The quality of annotation is further improved using taboo words that users are not allowed to type. Several variants of the ESP game have been developed, such as games for object region annotations [118, 138], video annotation [169], music annotation [8] and geographically-referenced photo annotation for landmark objects [10].

Another approach that motivates users is *crowdsourcing* that outsources problems performed by designated human (employee) to users on the Web [95]. In the field of multimedia annotation, one of the most famous crowdsourcing systems is Amazon's Mechanical Turk where anyone can post small tasks and specify prices paid for completing them [55].

Although large-scale annotation data can be obtained by the above Web-based systems, flexible retrieval like those presented in the previous section is difficult. This is because annotation has to be simple in order to maintain the usability. Also, one drawback of the GWAP approach is that users tend to maximise their scores, so collected labels only represent general properties of examples (e.g., colour and shape), but do not represent specifics or details [38]. Furthermore, it requires huge monetary cost to apply Mechanical Turk to large-scale data. To the best of our knowledge, no Web-based system that fully supports annotation of high-level meanings has been developed until now.

4.3 Interactive approaches

This section focuses on interactive approaches where users iteratively refine the retrieval performance based on the current result. These are needed because of the *user individuality*, which means that even for the same query different users may be interested in different data [165]. For example, for the query "horse", one user may look for shots showing "adult horse", while another may look for shots showing "child horse". In addition, it is often difficult for a user to precisely express his/her intent, because of the poor lexical vocabulary or the lack of proper positive examples. This is called the *intention gap* which is the discrepancy between user's search intent and the query specified by him/her [159]. Thus, the interactive refinement of retrieval results is necessary to overcome the user individuality and intention gap.

One of the most popular interactive approaches is *Relevance Feedback* (RF) that asks a user to provide feedback regarding the relevance or irrelevance of currently retrieved examples [165]. RF can be also considered as *active learning* that selects the most informative examples for improving the performance of a classifier, and asks the user to annotate them [139]. Using such RF or active learning, a classifier is refined so as to efficiently find desired examples. Rui et al. developed one of the earliest RF method that dynamically updates both feature weights and feature dimension weights based on the user-provided relevance score for each retrieved image [98]. Also, the most typical RF method uses an SVM

as a classifier where examples closest to its decision boundary are labelled by the user [128, 139]. This means that for such examples the prediction of the SVM is the most uncertain, so labelling them is useful for refining that SVM. In other words, examples far from the decision boundary are regarded to be reasonably classified, thus labelling them is redundant.

In [147], Wu and Zhang proposed an RF method using a random forest which predicts the relevance of an example to a query, by combining multiple tree classifiers built on different subsets of randomly sampled training examples. By refining the random forest based on additional training examples obtained by RF, multiple tree classifiers can cover the multimodal distribution of relevant examples in the feature space. This approach is extended to adaptive pattern discovery [148], which addresses RF by interactively discovering meaningful patterns of relevant examples. To facilitate pattern discovery, the authors present a dynamic feature extraction method, which aims to alleviate the curse of dimensionality by extracting a feature subspace using balanced information gain. The scientific achievements described above are integrated within the so called PatternQuest framework [149] that learns the patterns of interest (i.e., the distribution patterns of positive examples) using classification methods and RF.

The conventional RF requires a crisp binary decision to be made on the relevance of the retrieved images. However, user interpretation varies with respect to different information needs and perceptual subjectivity. In addition, users tend to learn from the retrieval results to further refine their information request. It is, therefore, inadequate to describe user's fuzzy perception of image similarity with crisp logic. In view of this, Yap et al. [152] proposed a 'fuzzy' RF approach which enables the user to make a fuzzy judgement for relevance ranking, whereas a radial basis function (RBF) network with local modelling structure is used for similarity learning. Another fuzzy RF method was described in [7], where a composite short-term and long-term learning approach is used to learn the semantics of an image. The short-term learning technique applies Fuzzy Support Vector Machine (FSVM) learning on user labelled and additional chosen image blocks to learn a more accurate boundary for separating the relevant and irrelevant blocks at each feedback iteration. The long-term learning technique applies a novel semantic clustering to adaptively learn and update the semantic concepts at each query session.

Apart from RF, interactive approaches include *browser-based* methods which provide user interfaces to facilitate finding relevant examples to a query [112]. Here, examples are organised based on some criteria. For example, Wilkins et al. developed the broadcast-based browsing interface [145]. This provides a user with a list of videos each of which contains highly-ranked shots (i.e., shots regarded as relevant by a classifier). Using video as the unit of retrieval result presentation, the user can explore the context where even if a highly-ranked shot is irrelevant, shots temporally close to it may be relevant. Snoek et al. devised the thread-based browsing interface to explore videos from multiple perspectives [116]. A thread presents a sequence of shots that are linked based on a certain type of similarity. The interface in [116] utilises four threads regarding the temporal similarity between shots, the visual similarity between shots, the visual similarity of shots to positive examples, and the history of explored shots. Recently, a live competition has been established where browser-based methods developed by different research groups are evaluated on the same experimental setting within view of the audience [103].

Another variant of interactive approaches is *visual query suggestion* which assists a user in precisely formulating a query by simultaneously suggesting keywords and images based on the initially provided keyword [159]. Compared to only suggesting keywords, adding images is useful for manifesting vague user's intent. In addition, the retrieval performance is improved by reranking images retrieved by keywords based on their visual similarities to

suggested images. In [159], the researchers developed a query suggestion mining method which firstly selects keywords that are statistically related and informative to the initially provided keyword. Then, images suggested with each selected keyword are obtained by clustering images associated with both of the initial and selected keywords, and choosing the representative ones. The above visual query suggestion is extended to object retrieval where given a query region of a certain object, regions showing the same object are retrieved from a large amount of images [39]. Considering the difficulty of specifying query regions effective for retrieval, the method suggests a query region of an object by guaranteeing that this object is also present in other images. To this end, repeated patterns of geometrically consistent local descriptors are firstly mined. Since such a pattern only represents a part of an object, bipartite clustering is performed to extract regions of the same object as a cluster of patterns that co-occur in the same images.

To sum up, although the performance is somehow improved by tuning features or classifiers based on RF, they are substantially the same and remain insufficient for representing high-level meanings. Browser-based methods just leave the most difficult task (i.e., interpretation of meanings) to humans. In addition, visual query suggestion just refines queries and is not related to the improvement of retrieval algorithms. Furthermore, considering the usability of interactive methods, user interaction should be done quickly. Thus, interactive methods are only useful for concepts which can be easily recognised from single images or shots.

5 LSMR based on human-machine cooperation

In this section, we review existing human-machine cooperation methods which incorporate knowledge about human interpretation into LSMR. As shown in Fig. 2, we sequentially present cognition-based, ontology-based and adaptive methods, and discuss future research topics for each type of method.

5.1 Cognition-based approaches

Cognitive science is an interdisciplinary study of mind and intelligence in order to theoretically explain how the human mind (thinking) works [85, 122, 127]. In particular, owing to psychological and neurological experiments, the human visual system is intensively investigated from the 'sensation' process which transducts the light (stimulus) received by the eye into neural signals, to the 'perception' process which translates neural signals into meanings. Since these processes are applicable to concept detection for deriving concepts from raw multimedia data, we define cognition-based methods as those that improve concept detection using knowledge about the human visual system. Below, as shown in Fig. 4, we review cognition-based methods by classifying them into two categories. The first includes methods using knowledge about concept meanings like hierarchical/spatio-temporal relations and attributes, and the second includes methods which aim to implement the human visual system.

5.1.1 Methods using knowledge about concept meanings

Hierachical relation A practical LSMR system is required to detect a variety of concepts. For example, over twenty-one thousand concepts are defined in ImageNet which is a huge concept vocabulary for images [25]. However, the conventional *one-vs-all* approach con-

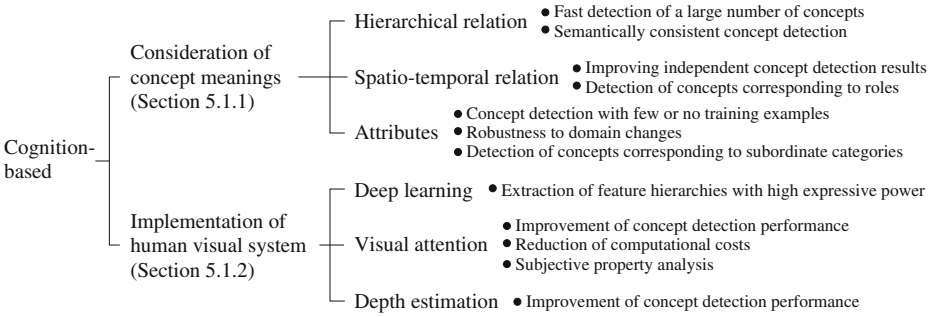

Fig. 4 Categorisation of cognition-based LSMR methods

structs a classifier for every concept, and tests all classifiers for each example. This is not suitable for treating a large number of concepts. Regarding this, when a human sees a new concept, he/she does not learn all of its appearance details, but just remembers its category and discriminative details [74]. This implies that the human forms a hierarchy of concepts. It is useful for not only detecting concepts quickly, but also achieving semantically consistent concept detection.

Using a large lexical ontology (WordNet) [31], Marszalek and Schmid developed a method which builds a classifier for each concept by defining training examples, so as to satisfy the hierarchical relation [74]. Here, positive examples are defined as the union of examples annotated with the concept's presence and those annotated with presences of its child concepts, whereas negative examples are the union of positive examples for its sibling concepts. Then, the top-down procedure is applied to a test example where the presence of a concept is examined only if its parent concept is detected. While the one-versus-all approach takes the computational complexity $O(n)$ where n is the number of concepts, the above method reduces it approximately to $O(n^{0.64})$ without loss of accuracy. Furthermore, Gao and Koller proposed a method which simultaneously constructs a hierarchy and classifiers by analysing distributions of positive examples for multiple concepts [35]. For each node in the hierarchy, a binary classifier is built by classifying concepts into two classes based on distributions of their positive examples. In particular, to maintain the generality of the classifier, the relaxed hierarchy structure is adopted where some concepts are flexibly ignored if their positive examples are difficult to classify. This method offers similar or slightly better performance with 2–5 times speed-up compared to the one-versus-all approach. Also, the top-down procedure propagates the error occurred at a concept for which the classifier is unreliable. To overcome this, Zhu et al. proposed an error recovery method which adjusts classifier's output for each concept by considering classifiers' outputs for its child and grandchild concepts [168]. Logistic regression is used to obtain the optimal weights, with which classifiers' outputs for the current, child and grandchild concepts are linearly combined to refine the detection of the current concept. The error recovery method improves the performance of the top-down procedure with 14–27 %, and saves the computational cost with 67.1–89.4 % compared to the one-versus-all approach.

Spatio-temporal relation Concepts do not appear in isolation. In other words, the presence of one concept can be a useful clue for detecting other concepts. Especially, concepts corresponding to roles in specific situations (e.g., *Athlete*) should be deduced based on relations to other concepts (e.g., *Playground* and *Running*). Below, we describe existing methods

that consider spatial relations among concepts within examples, as well as their temporal relations over shot sequences. Notice that the following discussion excludes methods for treating relations among features (or descriptors) like [106, 157], because such relations are not directly related to meanings due to the brittleness of features to changes in camera techniques and shooting environments.

Jiang et al. proposed a method which refines independent detection results of concepts by considering their correlations (co-occurrences) [48]. Using training examples, they first construct a graph where each node is a concept and the weight of an edge is defined as the Pearson product moment correlation between two concepts. Then, graph diffusion is performed to smooth detection results in each example, so that results for strongly correlating concepts become similar. This approach refines independent concept detection results by 11.8–15.6 %. Yi et al. developed another refinement method based on both spatial and temporal relations [153]. Regarding the former, they build a Conditional Random Field (CRF) which probabilistically estimates a refined detection result of a concept by considering independent detection results of this and correlated concepts. To treat temporal relations, another CRF is built to predict a refined detection result of a concept based on independent detection result of this and correlated concepts in surrounding shots. Experimental results showed that spatial and temporal relations respectively improve independent detection results by 20.6–36.4 % and 26.5–47.2 %, and their combination offers 29.9–53.1 % of improvement.

Furthermore, Chen et al. developed NEIL (Never Ending Image Learner) which continuously extracts visual knowledge (positive images and concept relations) from Internet scale data [21]. NEIL is based on semi-supervised learning. First, for each concept, seed images are collected through Google Image Search to built the initial classifier. Second, concept relations are extracted by computing co-occurrences based on classifiers' outputs. Third, NEIL selects additional positive images, each of which has large outputs of both the classifier for a concept and classifiers for its related concepts. Then, NEIL updates classifiers with additional positive images, and continuously repeats the second and third processes. As the result of running NEIL for 2.5 months, it could discover 400K positive examples and 1,700 concept relations for 2,237 concepts. It was also shown that using extracted relations improves the concept detection performance by 4.0–8.8 %.

Attributes A human can categorise even unseen concepts based on their characteristic appearances. For example, even though a human does not know the exact names of airplanes, he/she can discriminate between airplanes with and without propellers. In addition, without knowing the name *Zebra*, it can be distinguished from other horses based on the presence or absence of stripes. Such descriptions of concept appearances like parts (e.g., "propeller"), shapes (e.g., "round"), textures (e.g., "stripe") and non-verbal properties (e.g., "properties that dogs have but cats do not") are called *attributes* [30, 59]. These are semantically meaningful descriptions, and their automatic detection is relatively easy compared to concept detection. By representing each example using responses of attribute classifiers, concepts can be effectively detected with a small number of training examples, or only with plain text descriptions [30]. Also, as long as attribute classifiers are robust, the performance of attribute-based concept detection can be maintained irrespective of domain changes. Moreover, since attributes capture detailed appearances (e.g., propeller of an airplane) which are under-estimated by features based on many local descriptors, they are useful for detecting concepts corresponding to subordinate categories like *Jet_Airplane* and *Propeller_Airplane* of *Airplane* [19].

Farhadi et al. proposed a method which constructs an attribute classifier, where characteristic features are extracted by contrasting examples with and without the attribute within the

same category [30]. They showed that attribute-based concept detection can achieve similar performance only using 20 % of training examples, compared to feature-based detection. Lampert et al. verified the effectiveness of attributes for 'zero-shot learning' scenario. Here, no training examples are given, and a concept is detected using responses of attribute classifiers, and prior knowledge about which attributes are relevant to the concept [59]. Only for parts in attributes, one of the most popular methods is Deformable Part-based Model (DPM) which detects a region of a concept (object) by considering positions and deformations of parts, where each part is defined as a filter to capture its shape [32]. Compared to not-using parts, this method achieves 62.5 % and 33.3 % greater performance for localising *Person* and *Car*, respectively. Recently, Chai et al. demonstrated that concepts for subordinate categories can be effectively identified using features extracted from parts obtained by DPM (together with features from a region by image segmentation) [19]. Finally, since it is laborious to manually define a large vocabulary of attributes, Juneja et al. developed a method which automatically extracts distinctive attributes by firstly grouping regions of a certain concept into clusters, then examining the entropy of whether regions similar to those in each cluster are contained in different concepts [51]. Also, the method in [72] extracts attributes using crowdsourcing where users annotate corresponding regions in images showing the same concept, and differences in images displaying different concepts.

5.1.2 Implementing the human visual system

Deep learning The human brain processes visual information in hierarchical ways where neurons in the early visual areas extract simple features, which are transmitted to neurons in higher-level areas to form more complex features or concepts [57]. Inspired by this, *deep learning* has been developed to learn feature hierarchies with higher-level features formed by the composition of lower-level ones [11, 13]. This aims to construct multiple levels of feature representations where higher layers characterise more abstract features. Deep learning mainly offers the following three advantages (see [13] for more detail): The first is the expressive power where the combination of (distributed) features at each layer can define the exponential order of higher-level features, and this order is further exponentially increased by passing through layers. The second advantage is the invariance property where more abstract features are generally invariant to subtle changes in visual appearances. The last one is the explanatory factor that the learnt feature hierarchy can capture valuable patterns or structures underlying raw images or videos. Finally, a classifier for detecting a concept is created by using the learnt hierarchy as initialisation of a multi-layer neural network, or building a supervised classifier by constructing the feature vector of each example based on the hierarchy.

In [56], deep learning has been implemented as a multi-layer convolutional neural network which iteratively conducts convolution or pooling of outputs by neurons in the previous layer. Convolution works as feature extraction using filters each represented by weights of a neuron. On the other hand, pooling summarises outputs of neighbouring neurons to extract more abstract features. The multi-layer convolutional neural network is optimised by stochastic gradient descent which updates each weight of a neuron by back-propagating the derivative of training errors in terms of this weight. In ILSVRC 2012 which is a worldwide competition on large-scale image classification [43], the above mentioned method with the error rate of 15.3 % significantly outperformed the others (the second best error rate was 26.1 %). Also, Le et al. developed a nine-layer stacked sparse autoencoder to train concept detectors from unlabelled images [61]. Each layer consists of three sub-layers, filtering, pooling and normalisation, which respectively offer feature extraction from small

regions of the previous layer, the invariance of features (neighbouring neurons' outputs) to local deformation of visual appearances, and the range adjustment of features. The stacked sparse autoencoder is optimised layer-by-layer so that sparse features constructed at a layer can be accurately converted back into the ones at the previous layer. By training such a stacked autoencoder using 10 million unlabelled images with 16,000 cores, it was shown that the highest-level neurons characterise concepts like *Face*, *Cat_Face* and *Human_Body*. Moreover, compared to state-of-the-art methods, the multi-layer classifier using the stack autoencoder as the initialisation yields 15 % and 70 % performance improvement for 10,000 and 22,000 concept detection tasks, respectively.

Visual attention *Selective attention* is the brain's mechanism that determines which part of the sensory data is currently of the most interest [34]. This enables humans to conduct real-time decision-making by closely analysing selected parts in a large amount of data, captured by eyes and ears. *Visual attention* implements selective attention on images and videos to detect *salient regions* that are likely to attract users [34]. By filtering irrelevant regions, visual attention yields both the improvement of concept detection performance and the reduction of computational cost. In addition, visual attention can bridge the discrepancy between concept detection and retrieval. The former aims to detect concepts irrespective of various changing factors, while the latter needs to retrieve examples where concepts related to a query appear in regions drawing user's attention. For example, for the query "a car is moving", the user is not interested in an example where a *Car* moves in a small background region. Thus, visual attention facilitates analysing the subjective property of each example and achieving meaningful retrieval for humans.

Most of the visual attention methods analyse spatial distributions of biologically inspired features (e.g., brightness, contrast and curvature) in an example, and produce a *saliency map* which shows the degree of salience of each pixel [34]. Typically, pixels which are irregular compared to surrounding ones are regarded as salient. However, this kind of bottom-up approaches based only on features do not work well. Thus, researchers are exploring how to adopt top-down approaches using prior knowledge. One popular knowledge is 'contextual cueing' which means that a human can easily find a target object, when the visual context (i.e., spatial layout of objects) is similar to the past [34, 64]. Contextual cueing is implemented as supervised classification to build a classifier which detects salient regions in a test example, by referring to training examples where salient regions are annotated by manual or eye fixation records. In [64], multi-task learning is used where salient regions in examples with a particular visual context are detected by sharing classifiers among examples with correlated visual contexts. Compared to building a classifier for every visual context, each of the above classifiers is simultaneously built using more training examples with different visual contexts. This yields the better generalisation of the classifier. Also, based on *saccades* which are the transition of eye fixations (i.e., salient regions), Sugano et al. proposed an image segmentation method which performs joint clustering of fixation locations and seed regions [119]. This is formulated as energy minimisation on a graph, where each edge represents the cost for merging a pair of a fixation location and a seed region. Meaningful regions are extracted as the ones which are characterised by densely distributed fixations and uniform visual features. Experimental results showed that compared to a standard image segmentation method, jointly using fixations improves the performance with 17.5 to 50 %.

Depth estimation A 2D example (image or video frame) does not hold depth information in the real 3D world. As a result, for example, even though a *Person* stands in front of a *Table*, a 2D example shows that their regions are overlapping. In addition, despite the

fact that a *Person* throws a *Ball* far, their regions in a 2D example may be close to each other. Meanwhile, humans can easily recognise depth information in a 2D example. This has inspired researchers to develop methods that estimate depth information directly from 2D examples [52, 100]. To the best of our knowledge, there is no existing work which uses estimated depth information for concept detection. Nonetheless, its effectiveness is implied by concept detection using a depth sensor (typically Microsoft Kinect) [40]. For example, in [97], compared to only using 2D information, the accuracy of localising various concepts in indoor scenes is improved from 66.2 % to 71.4 % by additionally using depth information (please refer to [40] for other works).

In depth estimation, a classifier for predicting depth values in an example, is built using training examples where depth values are known with a depth sensor. Saxena et al. developed a method that firstly divides an example into 'superpixels' which are homogeneous small regions with similar properties [100]. They assume that each superpixel has the same depth value, and represent it with features useful for depth estimation. For example, a grass field viewed at a short distance has fine textures, while such textures are blurred when it is viewed at a large distance. In addition, parallel lines have larger variations in edge orientations as they are viewed at a more distant position. Using such texture and edge features, a Markov Random Field (MRF) is built to probabilistically estimate the depth value of each superpixel by considering its feature and relative depth values of nearby superpixels. Experimental results showed that depth values are reasonably estimated for 64.9 % of Web images. Karsh *el al.* developed a method which transfers depth values in training examples to a test example by assuming that, examples with similar meanings have similar spatial distributions of depth values [52]. Given a test example, the method firstly selects visually similar training examples. Then, depth values in each training example are transferred by extracting the correspondence of local regions between the training and test examples. In addition, the method can be applied to a video by smoothing depth values in each frame based on optical flows, and imposing moving objects to have similar depth values to the ground that they contact.

5.1.3 Discussion

Although Sections 5.1.1 and 5.1.2 present many existing cognition-based methods, we argue that they only use very limited knowledge about human visual system. Thus, much more knowledge needs to be adopted. One critical problem is that these methods use the same approach for every concept. Regarding this, machine learning methods are useful for concepts with basic categories, while attribute-based methods are effective for concepts with subordinate categories. Thus, one possible approach is to define the category of each concept, and then select a machine learning-based or attribute-based method depending on categories of concepts. In addition, some concepts significantly affect visual appearances of others, such as *Foggy*, *Nighttime* and *Dazzling*. Hence, modelling such relations seems valuable for selecting a classifier or modifying (transferring) an already trained classifier.

Also, deep learning seems a promising approach for implementing the human brain mechanism. However, current methods construct a feature hierarchy by just stacking the same type of layers of neurons, which only have a few variations of response functions. This hierarchy is much simpler than the human brain, where visual information encoded by the occipital part is divided into two interacting pathways, 'dorsal' and 'ventral', which are responsible for object categorisation and space/action analysis, respectively [57]. In addition, neurons have a variety of functionalities, so they cannot be optimised by the same parameter optimisation approach [57]. Thus, deep learning needs to be extended by adopting

a more complex hierarchy consisting of diverse types of neurons, based on a sophisticated optimisation method. Furthermore, visual attention and depth estimation can be considered to estimate information that cannot be directly observed from examples. In this context, estimating other information may be possible. For example, it is known that terahertz sensors produce waves which can pass through some objects (e.g., papers and plastic) to capture inside elements, and multispectral sensors can record invisible colour channels to characterise materials of objects. Thus, using these sensors' outputs as labels of training examples, it may be possible to build a classifier which can detect concepts (inside elements or materials) that humans infer from examples. Finally, although methods in Sections 5.1.1 and 5.1.2 have been developed independently, it seems beneficial to establish a framework for integrating these different categories of methods. We believe that their synergy will offer significant improvement of concept detection performance.

5.2 Ontology-based approaches

An ontology is a machine-readable representation of knowledge to explicitly specify concepts, properties of concepts and relations among concepts in a given domain [41]. We define ontology-based approaches as those that utilise ontologies to extract high-level meanings (i.e., events and contexts) based on detection results for multiple concepts. Figure 5 illustrates an overview where videos showing the event "birthday party" are identified. Note that although Fig. 5 uses video as the unit, it is straightforward to apply the following discussion to image or shot. In an ontology-based approach, every example is represented using *concept detection scores*, each representing a scoring value between 0 and 1 in terms of a concept's presence. A larger score indicates a higher likelihood of the concept's presence. In Fig. 5, as depicted by white-filled arrows, every example (video) is represented as a sequence of vectors each representing concept detection scores in a shot. Then, as indicated by black-filled arrows, based on this representation, a classifier is built using positive and negative examples, and used to discriminate between relevant and irrelevant test examples to the high-level meaning.

Since a high-level meaning is derived from the interaction among multiple concepts, the set of relevant examples has got a huge variance in the space of a low-level feature. Compared to this, owing to the recent progress, several concepts can be accurately detected. Thus, while each dimension in a low-level feature just represents physical values, detection scores for a concept (i.e., values in one dimension) represent appearances of a human-perceivable meaning. Hence, in the space of concept detection scores, the variance of relevant examples becomes smaller and can be modelled more easily. In other words, concept detection scores work as 'intermediate' features for a classifier to bridge between raw

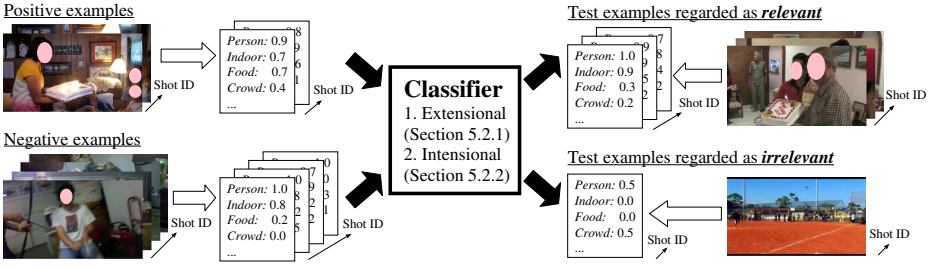

Fig. 5 An overview of an ontology-based approach

example representation and high-level meanings. Several publications reported the effectiveness of ontology-based approaches. For example, Tešič et al. showed that when using the same classifier (SVM), concept detection scores achieve 50–180 % higher event retrieval performance than colour and texture features [126]. In addition, Merler et al. reported that compared to high-dimensional features based on local descriptors (e.g., SIFT, HOG and HOF), concept detection scores yield the best performance [76]. In particular, the example representation using detection scores for 280 concepts only requires a 15 times smaller data space than high-dimensional features, where data sizes are crucial for the feasibility of LSMR. Furthermore, Mazloom et al. demonstrated that concept detection scores lead to 3.1–39.4 % performance improvement compared to the feature based on SIFT descriptors [75].

This section reviews existing ontology-based methods in terms of classifiers. As shown in the middle of Fig. 5, we categorise existing methods into *extensional* or *intensional*. The former includes methods that analyse training examples to extract concept relations useful for characterising high-level meanings. In other words, a high-level meaning is defined by providing its instances (i.e., training examples). On the other hand, intensional methods utilise knowledge about concept relations to characterise high-level meanings. That is, a high-level meaning is formed by its aspects (i.e., concept relations) known in advance. It should be noted that, in addition to classifiers, ontology-based approaches have two other important issues. The first is the construction of a concept vocabulary. For this, several large-scale concept vocabularies have recently become available, such as LSCOM (Large-Scale Concept Ontology for Multimedia) [80], ImageNet [25] and VSO (Visual Sentiment Ontology) [17]. The second issue is concept detection which is performed and improved by cognition-based methods in the previous section.

5.2.1 Extensional methods

In what follows, we classify extensional methods into two categories, where the one focuses on high-level meanings within images/shots, and the other targets those over shot sequences.

Within images/shots In general, methods in this category represent each example as a vector of concept detection scores, and build a classifier which discriminates between relevant and irrelevant examples to a high-level meaning. In other words, this classifier fuses detection scores for different concepts into a single *relevance score*, which indicates the relevance of an example to the meaning. Existing methods are roughly classified into four categories, *linear combination*, *discriminative*, *similarity-based* or *probabilistic*. Linear combination computes the relevance score of an example by weighting detection scores for multiple concepts. One popular weighting method is to use concept detection scores in positive examples. If the average detection score for a concept in positive examples is large, this concept is regarded as related to the query and associated with a large weight [82, 142]. Another popular method is text-based weighting where a concept is associated with a large weight if its name is lexically similar to a term in the text description of the query [82, 142]. The lexical similarity between a concept name and a term can be measured using a lexical ontology like WordNet. Discriminative methods construct a discriminative classifier (typically, SVM) using positive examples [82, 83]. The relevance score of an example is obtained as the classifier's output. Similarity-based methods compute the relevance score of an example as the similarity between positive examples and the example in terms of concept detection scores. The method in [63] uses the cosine similarity and a modified entropy as similarity measures. Probabilistic methods estimate a probabilistic distribution of concepts using detection

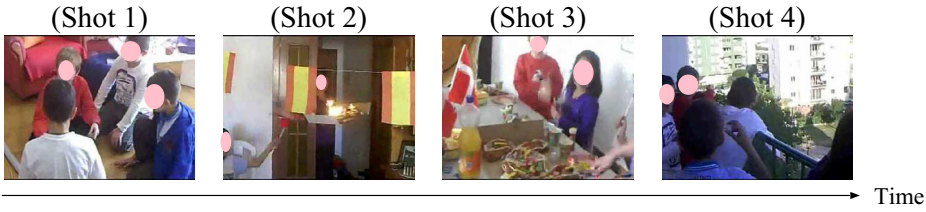

Fig. 6 An example video where the event "birthday party" is shown

scores in positive examples, and use it to compute the relevance score of an example. In [96], the relevance score of an image is computed as the similarity between the multinomial distribution of concepts estimated from positive examples and the one estimated from the image.

Over shot sequences When detecting a high-level meaning over shot sequences, one big problem is the difficulty of annotating the relevance of each shot. The reasons are two-fold: First, it is labour-intensive to annotate shots contained in each video. Second, due to the temporal continuity of meanings in a video, any segment can become a meaningful unit (see Section 4.1). Specifically, humans tend to relate each shot in a video to surrounding ones. Let us consider the video in Fig. 6 where the event "birthday party" is shown. There is no doubt that *Shot 2* and *3* show the birthday party, based on which *Shot 1* and *4* are related as chatting before the party and playing after it, respectively. This kind of shot relation makes it ambiguous to determine the boundary of a high-level meaning. In Fig. 6, one may think that the birthday party is shown only in *Shot 2* and *3*, while someone else may think that it is shown in the whole of the video, by regarding *Shot 1* and *4* as parts of the party. Thus, objective annotation is only possible at the video level in terms of whether each video contains a high-level meaning or not. A classifier needs to be build under this *weakly supervised setting*, where even if a training video is annotated as relevant to the meaning, it includes many irrelevant shots.

To overcome weakly supervised settings,[3] we developed an event detection method using a Hidden Conditional Random Field (HCRF) [109]. It is a probabilistic discriminative classifier with a set of hidden states. These states are used as the intermediate layer to discriminate between relevant and irrelevant shots to an event. Specifically, each shot in a video is assigned to a hidden state by considering its concept detection scores and transitions among hidden states. Such hidden states and transitions are optimised so as to maximise the discrimination between positive and negative videos. We showed that the optimised hidden states and transitions successfully capture concepts and their temporal relations, that are specific to the event. Sun and Nevatia proposed a method which extracts temporal concept transitions in an event using Fisher kernel encoding [120]. Using all training videos, they first build an HMM which works as a prior distribution, representing concept transitions in the general case. Then, the vector representation of a video is created by computing the difference between the actual transitions of concept detection scores in the video, and the transitions predicted by the HMM. Thereby, vectors of positive videos for an event represent

[3]Event detection under weakly supervised settings is being explored in TRECVID Multimedia Event Detection task [111]. Although some other methods (e.g., [124, 131]) can treat weakly supervised settings, they use low-level features, so are excluded from our discussion.

characteristic concept transitions by suppressing trivial transitions that are observed in many negative videos. Finally, to the best of our knowledge, Lu and Grauman developed the first metric which can quantify the context between two events, by finding concepts that appear in the first event and strongly influence the second one [70]. Such influences are measured by performing random walk on the bipartite graph, which consists of event and concept nodes. A concept is regarded as influential, if its ignorance leads to a dramatical decrease of the probability of transition between two event nodes. In [70], the above mentioned metric was used to create summaries consisting of events associated with semantically consistent contexts.

5.2.2 Intensional methods

Intensional methods exploit knowledge about concept relations to improve the detection performance of high-level meanings. To our best knowledge, existing methods can only deal with high-level meanings within examples. We will later discuss how to extend them for high-level meanings over shot sequences.

Deng et al. devised a method which computes the similarity between two examples based on the concept hierarchy [26]. The component similarity is computed as a weighted product between the detection score of an example for one concept, and the score of the counterpart example for another concept. Here, the weight is defined based on the lowest common ancestor of two concepts. For example, the component similarity for *Donkey* and *Horse* has a higher weight than the one for *Donkey* and *Keyboard*, because the common ancestor of the former concept pair *Equine* is more specific than that of the latter pair *Object*. The similarity between two examples is computed as the sum of component similarities. It was reported that using the concept hierarchy yields significant performance improvement [26].

Chen et al. proposed an interesting approach to organise independently detected concepts in an example using an ontology [20]. This ontology represents the hierarchy of concepts and their interactive relations. Based on this, the researchers devised an energy minimisation method which not only specialises detected concepts into more concrete ones, but also extracts their relations. The energy function consists of three terms: the first uses concept relations for estimating ontologically-consistent relations, the second term favours deep specialisation of concepts based on the concept hierarchy, and the last term uses pre-trained classifiers to examine the visual appearance of each specialised concept or estimated relation. By minimising this energy function, for instance, independently detected *Person* and *Ball* are specialised into *Basketball_Player* and *Basketball* which are linked with the relation of *Throw*. It was reported that compared to only using visual features, using the ontology reduces error rates of concept detection and relation estimation by 2.6–14.2 %.

Also, Guadarrama et al. developed a method which uses concept hierarchies obtained by text mining to extract events characterised by Subject/Verb/Object (SVO) relations [37]. First, SVO triplets are extracted from natural language descriptions of YouTube videos. Then, for each of S, V and O, concepts (i.e., subjects, verbs or objects) are clustered into a hierarchy based on their correlations. For each node (a set of concepts) in this hierarchy, a classifier is built where SVM's output is weighted based on both the specificity of the node and its WUP similarity to the other nodes. Given an example, such weighted outputs are used to select the best node for each of S, V and O. Finally, S, V or O is described by the WordNet concept with the highest sum of WUP similarities to concepts in the best node. Owing to the specificity and WUP similarity, the method can flexibly select not only specific concepts, but also concepts that are less specific but visually plausible.

5.2.3 Discussion

Compared to cognition-based methods in Section 5.1, much less ontology-based methods have been developed so far. Since concepts are just primitive meanings, the advancement of ontology-based methods is the most important topic to realise practical LSMR systems. Below, we point out three issues that deserve much research attention in the future.

1. **Uncertainties in concept detection:** Traditional ontology formalisms do not account for uncertainties, where an ontology itself is not uncertain. Compared to this, even using the most effective methods, it is still difficult to accurately detect any kind of concept. In particular, real-world examples are 'unconstrained' [50] in the sense that they can be taken by arbitrary camera techniques and in arbitrary shooting environments. Thus, it cannot be expected to detect concepts with 100 % of accuracy. Relying on such uncertain concept detection significantly damages the detection performance of high-level meanings. For this, we developed a method which handles uncertain concept detection based on *Dempster-Shafer Theory* (DST) [108, 110]. DST is a generalisation of Bayesian theory [27], where a probability is not assigned to a concept, but instead to a subset of concepts. This enables us to consider a probability that one of a set of concepts could be present in an example. By accumulating such probabilities for sets including a certain concept, we can define the *plausibility* which represents the upper bound probability of the concept's presence in the example. We incorporate such plausibilities into a probabilistic classifier, where plausibilities for a concept's presence are estimated as density ratios between positive and negative examples in terms of detection scores. That is, plausibilities are 'refined' detection scores by considering uncertainties in detecting the concept. Compared to directly using concept detection scores, using plausibilities yielded 19.1 % of performance improvement in detecting events within examples. We expect that such approaches for managing uncertain concept detection will be further explored to detect high-level meanings over shot sequences.
2. **Temporal continuity of a concept's presence:** Video editing is deemed as one main reason why intensional methods have not been developed for high-level meanings over shot sequences. The temporal order of shots can be easily distorted by inserting shots, that display different meanings than those of surrounding shots. Thus, it is difficult to reason a high-level meaning by specifying temporal positions of concepts. One possible solution is to model the temporal continuity of a concept's presence. Let us consider a sequence of three shots, where the first shot shows a *Person* bringing a *Birthday_Cake*, the second one shows *Person*s talking to each other, and the third one shows *Person*s eating the *Birthday_Cake*. Here, even if the *Birthday_Cake* is not shown in the second shot, humans can assume its existence. Hence, it is reasonable to consider that the *Birthday_Cake* is not absent in the second shot, but is just 'invisible'. To capture such a temporal continuity of a concept's presence, it seems effective to analyse appearance patterns of a concept related to the development of the story in a video. For instance, when a concept plays an important role, it is present in many shots, otherwise it is not. Based on this, our method in [104] can divide a video into shot sequences, characterised by probabilistically distinct patterns of the concept's presence. In such a shot sequence, the concept is assumed to be continuously present with the same degree of contribution to the story. This allows us to modify the detection score of a shot by considering scores of surrounding shots.

Then, reasoning of high-level meanings is performed on modified concept detection scores.

3. **Knowledge extraction:** To reason various high-level meanings, a large repository of concept relations is required. One lesson from the success of attributes, concepts, and curriculum learning in the next section, is to gradually increase levels of meanings. Thus, we need to first address the extraction of concept relations which characterise various events. This requires to solve the following three fundamental issues: The first is to define a standardised vocabulary of events. As described in [37], while vocabularies of concepts corresponding to nouns are already large-scale like LSCOM and ImageNet, existing works only focus on a handful of events. We expect that, as concepts in LSCOM have been defined through the collaboration of multimedia researchers, library scientists and end users [80], similar effort is needed to construct an event vocabulary. The second issue is how to examine concept relations in events. For temporal relations, it is crucial to model the temporal continuity of a concept's presence, described above. Similarly, depth estimation in Section 5.1.2 is vital to obtain semantically meaningful spatial relations among concepts. Assuming that these spatio-temporal concept relations could be successfully estimated, the last issue is to extract characteristic concept relations for events. Regarding this, in the next section, we will describe an efficient feedback approach to extract a large number of semantically meaningful concept relations with a small amount of user intervention. Finally, this feedback approach can be used to extract characteristic event relations for certain contexts.

5.3 Adaptive learning

Human learning can be considered as the repetition of the following process: Given a new problem, a human first monitors his/her performance, recognises a deficiency, and uses knowledge that he/she already owns to overcome the deficiency. By repeating this, the human can accumulate knowledge for solving diverse problems. *Metacognition* is a discipline to explore the process of how a human addresses a problem [3]. Assuming a cognitive system which simulates a functionality of the human mind, metacognition aims to monitor, model and control the behaviour of that system to effectively solve a problem. We position *adaptive learning* as metacognition for LSMR. Specifically, one LSMR method consists of various processes such as feature extraction, classifier construction, parameter tuning, training example collection, and so on. We define adaptive learning as methods which enhance or optimise one or more of the above mentioned processes based on knowledge about human learning. Below, we present two types of adaptive learning, *explanatory feedback* and *curriculum learning*.

1. **Explanatory feedback:** The traditional RF (or active learning) relies on the very restrictive communication between a classifier and a user, where the latter only informs the former of whether an example is relevant or irrelevant to a certain meaning (see Section 4.3). In the real world, a teacher makes much complex communication with a leaner. In particular, if the learner makes a mistake, the teacher tells him/her the reason for it.

   Based on this idea, Parkash and Parikh extended RF to *Explanatory Feedback* (EF), where if an example that a classifier selects as relevant to a meaning is judged to be irrelevant by a user, he/she can explain the reason for this mis-classification [89]. For

example, if an example showing *Forest* is mis-classified as *Street* by the classifier, the user can explain "this example is too natural to be a *Street*". EF is based on attributes with which an example is represented using responses of attribute classifiers (see Section 5.1.1). Since attributes are semantically meaningful descriptions, they can be used as a language between the classifier and the user to realise their complex communication. Especially, the irrelevance (negative) label assigned to the mis-classified example can be propagated to examples which contain the attribute explained by EF. In the above mentioned case, if examples have higher responses for the attribute "natural" than that of the mis-classified example, they are also regarded as irrelevant to *Street*. Like this, multiple negative examples are obtained only with one feedback, so that the classifier performance can be effectively improved. It was demonstrated that EF yields similar classifier performance only with one-fifth of iterations required in RF [89]. This method has been further improved by adopting the propagation of weighted irrelevance labels, on-the-fly update of attribute classifiers based on every feedback, and the example selection by estimating the entropy reduction resulting from the label propagation beginning at each example [16].

2. **Curriculum learning:** Human learning is highly organised based on a curriculum (education system), where children start to learn easier concepts and then build up more complex ones. Based on this, Bengio et al. proposed *curriculum learning* which builds a classifier by presenting training examples in a meaningful order, starting with easy examples and gradually introducing more difficult ones [12]. It should be noted that curriculum learning is not useful for a classifier with a convex optimisation function like SVM, because the global optimum can be found in any order of training examples. In other words, it aims to find a good local minimum to build a classifier with a non-convex optimisation function. In [12], a two-step curriculum was used for classification of shapes (rectangles, ellipses and triangles). Here, a multi-layer neural network is pre-trained using training examples with less variability in shape (squares, circles and equilateral triangles), and then trained using examples with diverse shape deformation. It was shown that compared to not-using pre-training, the two-step curriculum leads to about 23 % of performance improvement.

   Considering that curriculum learning in [12] relies on pre-defined 'easiness' values of examples (e.g., squares are easier to classify than general rectangles), Kumar et al. developed an iterative method where each iteration not only enlarges training examples by adding more difficult ones, but also updates the parameter of a classifier [58]. Assuming that labels of easy examples are easily predicted by the classifier, the researchers introduced binary variables each representing whether a training example is used to compute the optimisation function. Since the optimisation function favours correct classification of training examples, an easier example starts to be used from an earlier iteration. In [58], for object localisation using latent structural SVMs, the error rate 16.92 % was improved to 15.38 % by adopting the above curriculum learning. In a similar fashion, Ma et al. developed a video event detection method where a continuous-valued label, called 'fine-grained label', is assigned to each negative example, based on how easily it can be distinguished from positive examples [71]. Such fine-grained labels are initialised based on detection scores of negative examples for concepts, that are specific to positive ones. Then, the method jointly optimises fine-grained labels and two classifiers (one using concept detection scores and the other using features). That is, fine-grained labels are optimised so as to improve the performance of classifiers. The researchers reported that using fine-grained labels significantly enhances event detection [71].

5.3.1 Discussion

We consider EF as a promising approach to extract relations between high-level and low-level meanings, where these meanings are used as a language to make complex communication between humans and machines. Although the current EF only supports attribute-concept relations, it can be used to recursively extract concept-event and event-context relations (see the composition of meanings in Fig. 1). For the former relations, EF allows a user to explain concepts which caused the detection of an incorrect event for an example. In order for this EF to appropriately link concepts to the event, it is necessary to estimate meaningful spatio-temporal concept relations described in Section 5.2.3. Similarly, EF can be applied to event-context relations, where a falsely identified context is fixed by explaining the causative events. Also, easiness values of training examples in curriculum learning imply one interesting research topic to model *metalevel* features [3]. These do not characterise meanings, but capture aspects of features, classifiers and parameters in the LSMR pipeline. Apart from easiness values, for example, it is said that the performance of a decision tree can be estimated based on the number of nodes, depth, shape and so on [14]. By devising such metalevel features, we can decide or control a strategy to effectively utilise available features, classifiers and parameters for accurate retrieval.

6 Conclusion

In this paper, we reviewed existing LSMR methods including the ones that we developed. By tracing the history of machine-based and human-based methods, we stated that because of prioritising the generality and scalability for large-scale data, current methods lack knowledge about human interpretation which was used in classical syntactic or manual methods. Then, we presented existing human-machine cooperation methods which incorporate such knowledge into LSMR. In particular, we classified them into three types, cognition-based methods using knowledge about human visual system, ontology-based methods using knowledge about human inference, and adaptive learning methods using knowledge about human learning. Our retrospective survey indicates one remarkable difference between classical and human-machine cooperation methods. While the former just lists rules or templates as knowledge, the latter uses it to sophisticate computational models so as to maintain the generality and scalability. Since only limited knowledge is used in the current methods described in Section 5, we expect that much more knowledge will be adopted and integrated into LSMR.

To reach the aforementioned goal, our final suggestion is to consider the mutual relation among three types of human-machine cooperation methods. Figure 7 illustrates this relation. First, cognition-based methods are used to detect concepts, from which events and contexts are derived by ontology-based methods. In the opposite direction, concept relations that are used to characterise events and contexts in ontology-based methods, are useful for validating and refining concept detection results by cognition-based methods. In addition, detection of meanings (concepts, events and contexts) in these methods is enhanced by adaptive learning methods. Especially, EF provides means to effectively refine cognition-based and ontology-based methods by explaining reasons of false meaning detection. Meanwhile, the information (i.e., knowledge, features (concepts) and classifiers) of these methods is a material to extract metalevel features in adaptive learning methods, so that the latter can

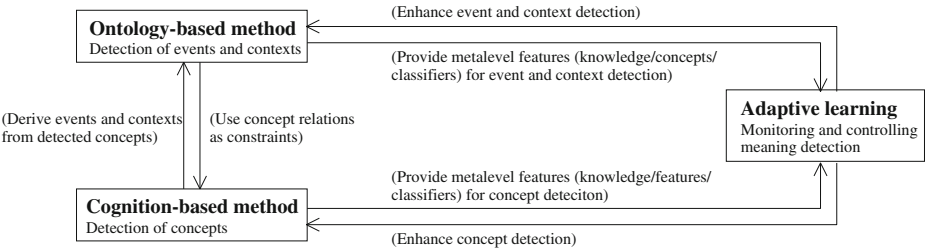

Fig. 7 An illustration of the relation among three types of human-machine cooperation methods (cognition-based, ontology-based and adaptive learning methods)

effectively control meaning detection in the former. Therefore, it is beneficial to develop a framework for unifying methods in the above mentioned mutually-related categories.

References

1. Adams B, Dorai C, Venkatesh S (2000) Novel approach to determining tempo and dramatic story sections in motion pictures. In: Proceedings of ICIP 2000, pp 283–286
2. Alham NK, Li M, Liu Y, Hammoud S (2011) A Map Reduce-based distributed SVM algorithm for automatic image annotation. Comput Math Appl 62(7):2801–2811
3. Anderson ML, Oates T (2007) A review of recent research in metareasoning and metalearning. AI Mag 28(1):7–16
4. Ando R, Shinoda K, Furui S, Mochizuki T (2006) Robust scene recognition using language models for scene contexts. In: Proceedings of MIR 2006, pp 99–106
5. Arandjelovic R, Zisserman A (2013) All about VLAD. In: Proceedings of CVPR 2013, pp 1578–1585
6. Ayache S, Quénot G (2008) Video corpus annotation using active learning. In: Proceedings of ECIR 2008, pp 187–198
7. Barrett S, Chang R, Qi X (2009) A fuzzy combined learning approach to content-based image retrieval. In: Proceedings of ICME 2009, pp 838–841
8. Barrington L, O'Malley D, Turnbull D, Lanckriet G (2009) User-centered design of a social game to tag music. In: Proceedings of HCOMP 2009, pp 7–10
9. Bay H, Tuytelaars T, Gool L (2006) SURF: speeded up robust features. In: Proceedings of ECCV 2006, pp 404–417
10. Bell M, Reeves S, Brown B, Sherwood S, MacMillan D, Ferguson J, Chalmers M (2009) EyeSpy: supporting navigation through play. In: Proceedings of CHI 2009, pp 123–132
11. Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127
12. Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of ICML 2009, pp 41–48
13. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
14. Bensusan H, Giraud-Carrier CG, Kennedy CJ (2000) A higher-order approach to meta-learning. In: Proceedings of ILP 2000
15. Bhatt C, Kankanhalli M (2011) Multimedia data mining: state of the art and challenges. Multimed Tools Appl 51(1):35–76
16. Biswas A, Parikh D (2013) Simultaneous active learning of classifiers & attributes via relative feedback. In: Proceedings of CVPR 2013, pp 644–651
17. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of MM 2013, pp 223–232
18. Catanzaro B, Sundaram N, Keutzer K (2008) Fast support vector machine training and classification on graphics processors. In: Proceedings of ICML 2008, pp 104–111
19. Chai Y, Lempitsky V, Zisserman A (2013) Symbiotic segmentation and part localization for fine-grained categorization. In: Proceedings of ICCV 2013, pp 321–328

20. Chen N, Zhou Q-Y, Prasanna V (2012) Understanding web images by object relation network. In: Proceedings of WWW 2012, pp 291–300
21. Chen X, Shrivastava A, Gupta A (2013) NEIL: extracting visual knowledge from web data. In: Proceedings of ICCV 2013, pp 1409–1416
22. Chu C et al (2007) Map-Reduce for machine learning on multicore. In: Schölkopf B, Platt J, Hoffman T (eds) NIPS 19. Birkhaüser, Cambridge, pp 281–288
23. Csurka G, Bray C, Dance C, Fan L (2004) Visual categorization with bags of keypoints. In: Proceedings of ECCV 2004 SLCV, pp 1–22
24. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):5:1–5:60
25. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of CVPR 2009, pp 248–255
26. Deng J, Berg A, Li FF (2011) Hierarchical semantic indexing for large scale image retrieval. In: Proceedings of CVPR 2011, pp 785–792
27. Denoeux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng 25(1):119–130
28. Djordjevic D, Izquierdo E, Grzegorzek M (2007) User driven systems to bridge the semantic gap. In: Proceedings of EUSIPCO 2007, pp 718–722
29. Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6:1889–1918
30. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Proceedings of CVPR 2009, pp 1778–1785
31. Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT Press, Cambridge
32. 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):1627–1645
33. François 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. Frintrop S, Rome E, Christensen HI (2010) Computational visual attention systems and their cognitive foundations: a survey. ACM Trans Appl Percept 7:6:1–6:39
35. Gao T, Koller D (2011) Discriminative learning of relaxed hierarchy for large-scale visual recognition. In: Proceedings of ICCV 2011, pp 2072–2079
36. Gemmell D, Vin H, Kandlur D, Venkat Rangan P, Rowe L (1995) Multimedia storage servers: a tutorial. IEEE Comput 28(5):40–49
37. Guadarrama S et al (2013) YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of ICCV 2013, pp 2712–2719
38. Gupta M, Li R, Yin Z, Han J (2010) Survey on social tagging techniques. SIGKDD Explor 12(1):58–72
39. Hamzaoui A, Letessier P, Joly A, Buisson O, Boujemaa N (2014) Object-based visual query suggestion. Multimed Tools Appl 68(2):429–454
40. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybern 43(5):1318–1334
41. Horridge M, Knublauch H, Rector A, Stevens R, Wroe C (2004) A practical guide to building OWL ontologies with the protege-OWL plugin, 1st edn. University of Manchester. http://home.skku.edu/samoh/class/sw/ProtegeOWLTutorial.pdf
42. Hsieh CJ, Chang KW, Lin CJ, Keerthi SS, Sundararajan S (2008) A dual coordinate descent method for large-scale linear svm. In: Proceedings of ICML 2008, pp 408–415
43. ImageNet Large Scale Visual Recognition Challenge (2012) (ILSVRC 2012). http://image-net.org/challenges/LSVRC/2012/index#workshop
44. Inoue N, Shinoda K (2012) A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors. IEEE Trans Multimed 14(4):1196–1205
45. Izquierdo E, Chandramouli K, Grzegorzek M, Piatrik T (2007) K-space content management and retrieval system. In: Proceedings of ICIAPW 2007, pp 131–136
46. Jain AK, Vailaya A, Wei X (1999) Query by video clip. Multimed Syst 7(5):369–384
47. Jégou H, Perronnin F, Douze M, Sánchez J, Perez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell 34(9):1704–1716
48. Jiang YG, Wang J, Chang SF, Ngo CW (2009) Domain adaptive semantic diffusion for large scale context-based video annotation. In: Proceedings of ICCV 2009, pp 1420–1427
49. Jiang YG, Yang J, Ngo CW, Hauptmann A (2010) Representations of keypoint-based semantic concept detection: a comprehensive study. IEEE Trans Multimed 12(1):42–53
50. Jiang YG, Bhattacharya S, Chang SF, Shah M (2013) High-level event recognition in unconstrained videos. Int J Multimed Inf Retr 2(2):73–101

51. Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: distinctive parts for scene classification. In: Proceedings of CVPR 2013, pp 923–930
52. Karsch K, Liu C, Kang S (2012) Depth extraction from video using non-parametric sampling. In: Proceedings of ECCV 2012, pp 775–788
53. Kashino K, Kurozumi T, Murase H (2003) A quick search method for audio and video signals based on histogram pruning. IEEE Trans Multimed 5(3):348–357
54. Kim YT, Chua TS (2005) Retrieval of news video using video sequence matching. In: Proceedings of MMM 2005, pp 68–75
55. Kittur A, Chi EH, Suh B (2008) Crowdsourcing user studies with mechanical turk. In: Proceedings of CHI 2008, pp 453–456
56. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: Bartlett P, Pereira F, Burges C, Bottou L, Weinberger K (eds) NIPS 25, pp 1106–1114
57. Krüger N et al (2013) Deep hierarchies in the primate visual cortex: what can we learn for computer vision? IEEE Trans Pattern Anal Mach Intell 35(8):1847–1871
58. Kumar MP, Packer B, Koller D (2010) Self-paced learning for latent variable models. In: Lafferty J, Williams CKI, Shawe-Taylor J, Zemel R, Culotta A (eds) NIPS 23, pp 1189–1197
59. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of CVPR 2009, pp 951–958
60. Lan T, Raptis M, Sigal L, Mori G (2013) From subcategories to visual composites: a multi-level framework for object detection. In: Proceedings of ICCV 2013, pp 369–376
61. Le Q, Ranzato M, Monga R, Devin M, Chen K, Corrado G, Dean J, Ng A (2012) Building high-level features using large scale unsupervised learning. In: Proceedings of ICML 2012
62. Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. ACM Trans Multimed Comput Commun Appl 2(1):1–19
63. Li X, Wang D, Li J, Zhang B (2007) Video search in concept subspace: a text-like paradigm. In: Proceedings of CIVR 2007, pp 603–610
64. Li J, Tian Y, Huang T, Gao W (2010) Probabilistic multi-task learning for visual saliency estimation in video. Int J Comput Vis 90(2):150–165
65. Lin CY, Tseng BL, Smith JR (2003) Video collaborative annotation forum: establishing ground-truth labels on large multimedia datasets. In: Proceedings of TRECVID 2003
66. Litayem S, Joly A, Boujemaa N (2012) Hash-based support vector machines approximation for large scale prediction. In: Proceedings of BMVC 2012, pp 86.1–86.11
67. Liu X, Zhuang Y, Pan Y (1999) A new approach to retrieve video by example video clip. In: Proceedings of MM 1999, pp 41–44
68. Liu Y, Zhang D, Lu G, Ma W (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recognit 40(1):262–282
69. Lowe D (1999) Object recognition from local scale-invariant features. In: Proceedings of ICCV 1999, pp 1150–1157
70. Lu Z, Grauman K (2013) Story-driven summarization for egocentric video. In: Proceedings of CVPR 2013, pp 2714–2721
71. Ma Z, Yang Y, Xu Z, Sebe N, Hauptmann AG (2013) We are not equally negative: fine-grained labeling for multimedia event detection. In: Proceedings of MM 2013, pp 293–302
72. Maji S, Shakhnarovich G (2014) Part and attribute discovery from relative annotations. Int J Comput Vis 108(1–2):82–96
73. Maji S, Berg A, Malik J (2008) Classification using intersection kernel support vector machines is efficient. In: Proceedings of CVPR 2008, pp 1–8
74. Marszalek M, Schmid C (2007) Semantic hierarchies for visual object recognition. In: Proceedings of CVPR 2007, pp 1–7
75. Mazloom M, Habibian A, Snoek CG (2013) Querying for video events by semantic signatures from few examples. In: Proceedings of MM 2013, pp 609–612
76. Merler M, Huang B, Xie L, Hua G, Natsev A (2012) Semantic model vectors for complex video event recognition. IEEE Trans Multimed 14(1):88–101
77. Monaco J (1981) How to read a film. Oxford University Press, Oxford
78. Nam J, Alghoniemy M, Tewfik A (1998) Audio-visual content-based violent scene characterization. In: Proceedings of ICIP 98, pp 353–357
79. Naphade MR, Smith JR (2004) On the detection of semantic concepts at TRECVID. In: Proceedings of MM 2004, pp 660–667
80. Naphade M, Smith J, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91

81. Natsev AP, Naphade MR, Tešić J (2005) Learning the semantics of multimedia queries and concepts from a small number of examples. In: Proceedings of MM 2005, pp 598–607
82. Natsev AP, Haubold A, Tešić J, Xie L, Yan R (2007) Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: Proceedings of MM 2007, pp 991–1000
83. Ngo C et al (2009) VIREO/DVM at TRECVID 2009: high-level feature extraction, automatic video search and content-based copy detection. In: Proceedings of TRECVID 2009, pp 415–432
84. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: Proceedings of ECCV 2006, pp 490–503
85. Ogiela M, Tadeusiewicz R (2010) Towards new classes of cognitive vision systems. In: Proceedings of CISIS 2010, pp 851–855
86. Oh J, Bandi B (2002) Multimedia data mining framework for raw video sequences. In: Proceedings MDM/KDD 2002, pp 23–26
87. Oomoto E, Tanaka K (1993) OVID: design and implementation of a video-object database system. IEEE Trans Knowl Data Eng 5(4):629–643
88. Pan JY, Faloutsos C (2001) VideoGraph: a new tool for video mining and classification. In: Proceedings of JCDL 2001, pp 116–117
89. Parkash A, Parikh D (2012) Attributes for classifier feedback. In: Proceedings of ECCV 2012, pp 354–368
90. PASCAL Visual Object Classes. http://pascallin.ecs.soton.ac.uk/challenges/VOC/
91. Pattanasri N, Chatvichienchai S, Tanaka K (2005) Towards a unified framework for context-preserving video retrieval and summarization. In: Proceedings of ICADL 2005, pp 119–128
92. Peng Y, Ngo CW (2005) EMD-based video clip retrieval by many-to-many matching. In: Proceedings of CIVR 2005, pp 71–81
93. Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: Proceedings of CVPR 2007, pp 1–8
94. Petkovic M, Jonker W (2002) Content-based video retrieval: a database perspective. Kluwer Academic Publishers, Norwell
95. Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of CHI 2011, pp 1403–1412
96. Rasiwasia N, Moreno P, Vasconcelos N (2007) Bridging the gap: query by semantic example. IEEE Trans Multimed 9(5):923–938
97. Ren X, Bo L, Fox D (2012) RGB-(D) scene labeling: features and algorithms. In: Proceedings of CVPR 2012, pp 2759–2766
98. Rui Y, Huang T, Ortega M, Mehrotra S (1998) Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Trans Circuits Syst Video Technol 8(5):644–655
99. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 77(1-3):157–173
100. Saxena A, Sun M, Ng AY (2009) Make3D: learning 3D scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell 31(5):824–840
101. Scherp A, Mezaris V (2014) Survey on modeling and indexing events in multimedia. Multimed Tools Appl 70(1):7–23
102. Schmid C, Mohr R (1997) Local grayvalue invariants for image retrieval. IEEE Trans Pattern Anal Mach Intell 19(5):530–535
103. Schoeffmann K et al (2014) The video browser showdown: a live evaluation of interactive video search tools. Int J Multimed Inf Retr 3(2):113–127
104. Shirahama K, Uehara K (2008) A novel topic extraction method based on bursts in video streams. Int J Hybrid Inf Technol 1(3):21–32
105. Shirahama K, Uehara K (2012) Kobe university and Muroran institute of technology at TRECVID 2012 semantic indexing task. In: Proceedings of TRECVID 2012, pp 239–247
106. Shirahama K, Ideno K, Uehara K (2007) A time-constrained sequential pattern mining for extracting semantic events in videos. In: Petrushin V, Khan L (eds) Multimedia data mining and knowledge discovery. Springer, London, pp 404–426
107. Shirahama K, Matsuoka Y, Uehara K (2012) Event retrieval in video archives using rough set theory and partially supervised learning. Multimed Tools Appl 57(1):145–173
108. Shirahama K, Kumabuchi K, Uehara K (2013) Video retrieval by learning uncertainties in concept detection from imbalanced annotation data. In: Proceedings of MMEDIA 2013, pp 19–24
109. Shirahama K, Grzegorzek M, Uehara K (2014) Multimedia event detection using hidden conditional random fields. In: Proceedings of ICMR 2014, pp 9:9–9:16

110. Shirahama K, Kumabuchi K, Grzegorzek M, Uehara K (2014) Video retrieval based on uncertain concept detection using dempster-shafer theory. In: Baughman AK, Gao J, Pan JY, Petrushin V (eds) Multimedia data mining and analytics: disruptive innovation. Springer, London
111. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proceedings of MIR 2006, pp 321–330
112. Smeaton AF, Wilkins P, Worring M, de Rooij O, Chua TS, Luan H (2008) Content-based video retrieval: three example systems from TRECVid. Int J Imaging Syst Technol 18(2–3):195–201
113. Smeulders A, 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
114. Snoek CGM, Worring M (2009) Concept-based video retrieval. Found Trends Inf Retr 2(4):215–322
115. Snoek CGM, Worring M, Geusebroek JM, Koelma D, Seinstra F (2005) On the surplus value of semantic video analysis beyond the key frame. In: Proceedings of ICME 2005, pp 386–389
116. Snoek C et al (2009) The MediaMill TRECVID 2009 semantic video search engine. In: Proceedings of TRECVID 2009, pp 226–238
117. Staab S, Scherp A, Arndt R, Troncy R, Grzegorzek M, Saathoff C, Schenk S, Hardman L (2008) Semantic multimedia. In: Baroglio C, Bonatti PA, Maluszynski J, Marchiori M, Polleres A, Schaffert S (eds) Reasoning web, chap 4. Springer LNCS 5224, San Servolo, pp 125–170
118. Steggink J, Snoek C (2011) Adding semantics to image-region annotations with the name-it-game. Multimed Syst 17(5):367–378
119. Sugano Y, Matsushita Y, Sato Y (2013) Graph-based joint clustering of fixations and visual entities. ACM Trans Appl Percept 10(2):10:1–10:16
120. Sun C, Nevatia R (2013) ACTIVE: activity concept transitions in video event classification. In: Proceedings of ICCV 2013, pp 913–920
121. Tadeusiewicz R (2007) Intelligent web mining for semantically adequate images. In: Proceedings of AWIC 2007, pp 3–10
122. Tadeusiewicz R (2007) What does it means automatic understanding of the images? In: Proceedings of IST 2007, pp 1–3
123. Tanaka K, Ariki Y, Uehara K (1999) Organization and retrieval of video data (special issue on new generation database technologies). IEICE Trans Inf Syst 82(1):34–44
124. Tang K, Fei-Fei L, Koller D (2012) Learning latent temporal structure for complex event detection. In: Proceedings of CVPR 2012, pp 1250–1257
125. Tao D, Tang X, Li X, Wu X (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans Pattern Anal Mach Intell 28(7):1088–1099
126. Tešič J, Natsev AP, Smith JR (2007) Cluster-based data modeling for semantic video search. In: Proceedings of CIVR 2007, pp 595–602
127. Thagard P (2007) Cognitive science. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/fall2008/entries/cognitive-science/
128. Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: Proceedings of MM 2001, pp 107–118
129. Torralba A., Fergus R., Freeman W. (2008) 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach Intell 30(11):1958–1970
130. Uehara K, Oe M, Maehara K (1996) Knowledge representation, concept acquisition and retrieval of video data. In: Proceedings of CODAS 1996, pp 527–534
131. Vahdat A, Cannons K, Mori G, Oh S, Kim I (2013) Compositional models for video event detection: a multiple kernel learning latent variable approach. In: Proceedings of ICCV 2013, pp 1185–1192
132. 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
133. van de Sande KEA, Gevers T, Snoek CGM (2011) Empowering visual categorization with the GPU. IEEE Trans Multimed 13(1):60–70
134. Vapnik V (1998) Statistical learning theory. Wiley-Interscience
135. Volkmer T, Smith JR, Natsev AP (2005) A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In: Proceedings of MM 2005, pp 892–901
136. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of CHI 2004, pp 319–326
137. von Ahn L, Dabbish L (2008) Designing games with a purpose. Commun ACM 51(8):58–67
138. von Ahn L, Liu R, Blum M (2006) Peekaboom: a game for locating objects in images. In: Proceedings of CHI 2006, pp 55–64
139. Wang M, Hua XS (2011) Active learning in multimedia annotation and retrieval: a survey. ACM Trans Intell Syst Technol 2(2):10:1–10:21

140. Wang XJ, Zhang L, Liu M, Li Y, Ma WY (2010) ARISTA—image search to annotation on billions of web photos. In: Proceedings of CVPR 2010, pp 2987–2994
141. Wang H, Klaser A, Schmid C, Liu CL (2011) Action recognition by dense trajectories. In: Proceedings of CVPR 2011, pp 3169–3176
142. Wei XY, Jiang YG, Ngo CW (2011) Concept-driven multi-modality fusion for video search. IEEE Trans Circuits Syst Video Technol 21(1):62–73
143. Weiss R, Duda A, Gifford D (1994) Content-based access to algebraic video. In: Proceedings of ICMCS 1994, pp 140–151
144. Westermann U, Jain R (2007) Toward a common event model for multimedia applications. IEEE Multimed 14(1):19–29
145. Wilkins P et al (2007) K-space at TRECVid 2007. In: Proceedings of TRECVID 2007
146. Woelk D, Kim W, Luther W (1986) An object-oriented approach to multimedia databases. In: Proceedings of SIGMOD 1986, pp 311–325
147. Wu Y, Zhang A (2003) An adaptive classification method for multimedia retrieval. In: Proceedings of ICME 2003, pp 757–760
148. Wu Y, Zhang A (2003) Adaptive pattern discovery for interactive multimedia retrieval. In: Proceedings of CVPR 2003, pp 649–655
149. Wu Y, Zhang A (2004) PatternQuest: learning patterns of interest using relevance feedback in multimedia information retrieval. In: Proceedings of ICME 2004, pp 261–264
150. 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 LS-MMRM 2009, pp 35–42
151. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of CVPR 2009, pp 1794–1801
152. Yap KH, Wu K (2003) Fuzzy relevance feedback in content-based image retrieval. In: Proceedings of ICICS-PCM 2003, pp 1595–1599
153. Yi J, Peng Y, Xiao J (2013) Exploiting semantic and visual context for effective video annotation. IEEE Trans Multimed 15(6):1400–1414
154. Yoshitaka A, Ishii T, Hirakawa M, Ichikawa T (1997) Content-based retrieval of video data by the grammar of film. In: Proceedings of VL 1997, pp 310–317
155. Yu K, Zhang T, Gong Y (2009) Nonlinear learning using local coordinate coding. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (eds) NIPS 22, pp 2223–2231
156. Yuan J, Tian Q, Ranganath S (2004) Fast and robust search method for short video clips from large video collection. In: Proceedings of ICPR 2004, pp 866–869
157. Yuan J, Wu Y, Yang M (2007) Discovery of collocation patterns: from visual words to visual phrases. In: Proceedings of CVPR 2007, pp 1–8
158. Zettsu K, Uehara K, Tanaka K, Kimura N (1997) A time-stamped authoring graph for video databases. In: Proceedings of DEXA 1997, pp 192–201
159. Zha ZJ, Yang L, Mei T, Wang M, Wang Z, Chua TS, Hua XS (2010) Visual query suggestion: towards capturing user intent in internet image search. ACM Trans Multimed Comput Commun Appl 6(3):13:1–13:19
160. Zhai Y, Rasheed Z, Shah M (2004) A framework for semantic classification of scenes using finite state machines. In: Proceedings of CIVR 2004, pp 279–288
161. Zhai Y, Yilmaz A, Shah M (2005) Story segmentation in news videos using visual and text cues. In: Proceedings of CIVR 2005, pp 92–102
162. Zhang H, Gong Y, Smoliar S, Yeo Tan S (1994) Automatic parsing of news video. In: Proceedings of ICMCS 1994, pp 45–54
163. 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 Vis 73(2):213–238
164. Zhong D, Chang SF (2001) Structure analysis of sports video using domain models. In: Proceedings of ICME 2001, pp 713–716
165. Zhou XS, Huang TS (2003) Relevance feedback in image retrieval: a comprehensive review. Multimed Syst 8(6):536–544
166. Zhou H, Kimber D (2006) Unusual event detection via multi-camera video mining. In: Proceedings ICPR 2006, pp 1161–1166
167. Zhu X, Wu X, Elmagarmid AK, Feng Z, Wu L (2005) Video data mining: semantic indexing and event detection from the association perspective. IEEE Trans Knowl Data Eng 17(5):665–677
168. Zhu S, Wei XY, Ngo CW (2013) Error recovered hierarchical classification. In: Proceedings of MM 2013, pp 697–700
169. Zwol RV, Garcia L, Ramirez G, Sigurbjornsson B, Labad M (2008) Video tag game. In: Proceedings of WWW 2008

Kimiaki Shirahama received his B.E., M.E. and Ph.D. degrees in Engineering from Kobe University, Japan in 2003, 2005 and 2011, respectively. Currently, through the postdoctral fellowship of Japan Society for the Promotion of Science (JSPS), he is a postdoctral researcher at the Research Group for Pattern Recognition in University of Siegen, Germany. He is also an assistant professor in College of Information and Systems at Muroran Institute of Technology, Japan. His research interests include multimedia data processing, machine learning, data mining and virtual reality. He is a member of ACM SIGKDD, ACM SIGMM, the Institute of Image Information and Television Engineers in Japan (ITE), Information Processing Society of Japan (IPSJ) and the Institute of Electronics, Information and Communication Engineering in Japan (IEICE).

Marcin Grzegorzek is Assistant Professor at the University of Siegen heading the Research Group for Pattern Recognition in the Institute for Vision and Graphics. Currently, he is also Fellow of the Think Tank "Stiftung Neue Verantwortung" in Berlin where he leads the project "Cognitive Robotics". Furthermore, he is Principal Investigator in the Research Training Group 1564 "Imaging New Modalities". Marcin received his Master's Degree from the Silesian University of Technology in Gliwice in 2002 and PhD with distinction from the University of Erlangen-Nuremberg in 2007. From 2006 to 2008 he was Research Assistant at the Queen Mary, University of London. From 2008 to 2010 he worked as Lecturer for the University of Koblenz-Landau. Marcin was General Chair of the SAMT 2010, the 5th International Conference on Semantic and Digital Media Technologies. Moreover, he is Guest Editor of the MTAP (Multimedia Tools and Applications) Journal and acts as a Secretary in the Executive Board of the SMaRT (Semantic Multimedia Research and Technology) Association. His research interests include multimodal object recognition and scene analysis, semantic multimedia analysis and retrieval, as well as behavioural biometry and medical image processing. He is author and coauthor of more than 60 papers.