Chapter 1
Multimedia Semantics

In recent years, the production, storage, and sharing of the exponentially increasing number of multimedia resources became simple, owing to lower hardware costs, new web standards, and free hosting options on social media and video sharing portals. However, the contents of multimedia resources are, for the most part, meaningless to software agents, preventing automated processing, which is very much desired not only in multimedia retrieval, but also in machine vision. There are multiple approaches to address this issue, such as the machine-readable annotation of the depicted concepts, the formal description of scenes, and machine learning with pretrained classifiers, the latter of which is the primary means of automated structured multimedia annotation.

1.1 Rationale

So far, the majority of multimedia semantics research has focused on image understanding, and to a far lesser extent on audio and video semantics. Since the term *multimedia* refers to various combinations of two or more content forms, including text, audio, image, animation, video, and interactive content, images alone are not multimedia contents, but can be components of multimedia contents. Yet, the literature often uses the term multimedia to techniques and tools that are limited to capturing image semantics.

Inherently, video interpretation is far more complex than image understanding. Knowledge acquisition in video content analysis involves the extraction of spatial and temporal information, which can be used for a wide range of applications including, but not limited to, face recognition, object tracking, dynamic masking, tamper detection, abandoned luggage detection, automated number plate recognition, and lane departure warning. Without context and *semantics* (meaning), however, even basic tasks are limited or infeasible. For example, an attack cannot be differentiated from self-defense or training without seeing the preceding events. As

L.F. Sikos, *Description Logics in Multimedia Reasoning*,
DOI 10.1007/978-3-319-54066-5_1

a consequence, video event identification alone is insufficient for classification. Automated scene interpretation and video understanding rely on the formal representation of human knowledge, which is suitable for automated subtitle generation, intelligent medical decision support, and so on. While computers can be trained to recognize features based on signal processing, they cannot interpret sophisticated visual contents without additional semantics (see Table 1.1).

Table 1.1 Major differences between human understanding and computer interpretation of video scenes

Humans	Computers
Intelligence	
Real-time understanding is straightforward in most cases, although movies with a complex plot may need to be watched again to be fully understood	Overwhelming amount of information to process; algorithms and methods are often insufficient, making video understanding infeasible even without time constraints, yet alone in near-real time or real time
Context is understood from plot, title, events, genre, etc.	Potential interpretations are extremely confusing; metadata, if available, can be combined with concepts mapped to common sense knowledge bases or ontologies
Visual content is understood (even without colors)	Automatically extractable features and their statistics convey no information about the actual visual content (nothing is self-explanatory)
Years or decades of life experience and learning make it possible to recognize virtually anything	Training from a few hundred or thousand clips provides a very limited recognition capability
General understanding of how the universe works (e.g., common sense, naïve physics)	Only tiny, isolated representations of the world are formalized, therefore unconstrained video scenes cannot be processed efficiently
Understanding of human behavior enables prediction	Only fractions of human behavior are encoded, so software agents cannot expect upcoming movements
The human mind and eyes are adaptive and recognize persons or objects moving to or in darkness, or in noisy or less detailed recordings (e.g., old VHS video, small-resolution video)	If the noise-signal ratio falls below a threshold, algorithms perform poorly
Spatial information	
3D projected to 2D can be interpreted by stereovision: planes of graphical projection with multiple vanishing points are understood, which enables perspective viewing	Most videos have no depth information, although proprietary and standardized 3D recording and playback mechanisms are available; RGB-D and Kinect depth sensors can provide depth information
3D objects are recognized from most angles	Training provides information for particular viewing angles only—recognition from different viewpoints is problematic; scale-/rotation-invariant features are used for object tracking in videos
Partially covered objects and persons are relatively easily recognized	Occlusion is problematic

(continued)

Table 1.1 (continued)

Humans	Computers
Temporal information	
Continuity; events and happenings are understood even in the case of nonlinear narratives with extensive flashbacks and flashforwards (although movies with a very complex plot might be watched again to be fully understood)	Very few mechanisms for complex event detection; videos are usually compressed using lossy compression, therefore only certain frames can be used; no information can be obtained on complex events from signal processing
Information fusion	
Seamless/straightforward understanding of simultaneous multimodal information playback (e.g., video with audio and subtitle(s), hypervideo)	Information fusion is desired, which needs more research
Audio channel is understood relatively easily and conveys additional information	Algorithms for detecting distinct noises (e.g., gunshot, screaming) are available; complex audio analysis is a challenge
Subtitles and closed captions can be read by most humans and convey additional information	Text-based, timestamped subtitle files can be processed very efficiently; however, incorporating the obtained information into higher-level video understanding is still challenging

1.2 Feature Extraction and Feature Statistics for Classification

Multimedia features extracted from media resources can be converted into numerical or symbolic form, which enables the automated processing of core characteristics. Low-level features, which capture the perceptual saliency of media signals, include visual features (e.g., color, texture, shape), audio features (e.g., loudness, pitch, timber), and text features (e.g., speaking rate and pause length calculated by processing closed captions). Combining the results of video, audio, and subtitle analysis often provides complementary information. Production and licensing metadata, when available, can be used for aggregated semantic enrichment of media resources.

A wide range of well-established algorithms exists for automatically extracting low-level video features, as, for example, fast color quantization to extract the dominant colors [1] or Gabor filter banks to extract homogeneous texture descriptors [2]. Some of these features, such as motion vectors, are employed by video compression algorithms of state-of-the-art video codecs, such as H.264/MPEG-4 AVC and H.265/HEVC, and by video analysis.

The state-of-the-art video classification approaches exploit sparse local keypoint features, i.e., salient patches that contain rich local information about an image or a video frame. They apply the *bag of visual words* (BoVW) model using local aggregated visual descriptors, typically histogram of oriented gradients (HOG), histogram of optical flow (HOF), or motion boundary histograms (MBH).

Histograms are based on accumulative statistics that are not affected by small local changes of the content and are invariant to common transformations, such as signal scaling or coordinate shift. Well-established algorithms utilizing such descriptors are efficient in classification, video clip matching, and object recognition, but not necessarily in video understanding, particularly when it comes to scene interpretation and knowledge discovery. The bag-of-words model applies a visual vocabulary generated by grouping similar keypoints into a large number of clusters and handling each cluster as a visual word. A histogram of visual words can be constructed by mapping the keypoints back into the vocabulary, which provides the feature clue for multimedia indexing and classification.

1.3 Machine Learning for Multimedia Understanding

Low-level multimedia descriptors are typically fed into a recognition system powered by supervised learning, such as SVM classifiers (support vector machines), which look for an optimal hyperplane to find the most probable interpretation, such as via a Lagrangian optimization problem [3] (see Eq. 1.1):

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}\|\beta\|^2 \quad \text{s.t.} y_i\left(\beta^{\mathrm{T}} x_i + \beta_0\right) \geqslant 1 \text{ for } \forall i \tag{1.1}$$

where $\beta^{\mathrm{T}} x_i + \beta_0$ represents a hyperplane, β the weight vector of the optimal hyperplane, β_0 the bias of the optimal hyperplane, y_i the labels of the training examples, and x_i the training examples. Since there is an infinite number of potential representations of the optimal hyperplane, there is a convention to choose the one where $\beta^{\mathrm{T}} x_i + \beta_0 = 1$. Once a classifier is trained on images depicting the object of interest (positive examples) and on images that do not (negative examples), it can make decisions regarding the probability of object match in other images (i.e., object recognition). Complex events of unconstrained real-world videos can also be efficiently detected and modeled using an intermediate level of semantic representation using support vector machines [4].

Bayesian networks are suitable for content-based semantic image understanding by integrating low-level features and high-level semantics [5]. By using a set of images for training to derive simple statistics for the conditional probabilities, Bayesian networks usually provide more relevant concepts than discriminant-based systems, such as neural networks. Papadopoulos et al. proposed a machine learning approach to image classification, which combines global image classification and local, region-based image classification through information fusion [6].

Other machine learning models used for multimedia semantics include hidden Markov models (HMM), k-nearest neighbor (kNN), Gaussian mixture models (GMM), logistic regression, and Adaboost.

1.4 Object Detection and Recognition

The research interest for machine learning applications in object recognition covers still images, image sequences [7], and videos, in which spatiotemporal data needs to be taken into account for machine interpretation. There are advanced algorithms for video content analysis, such as the Viola-Jones and Lienhart-Maydt object detection algorithms [8, 9], as well as the SIFT [10], SURF [11], and ORB [12] keypoint detection algorithms. The corresponding descriptors can be used as positive and negative examples in machine learning, such as support vector machines and Bayesian networks, for keyframe analysis and, to a lesser extent, video scene understanding. Beyond the general object detection algorithms, there are algorithms specifically designed for human recognition. Persons can be recognized by, among others, shape, geometric features such as face [13], and behavioral patterns such as gait [14].

Based on the detected or recognized objects, machine learning can be utilized via *training*, which relies on a set of training samples (*training dataset*). For example, by creating a training dataset containing cropped, scaled, and eye-aligned images about different facial expressions of a person, the face of the person can be automatically recognized in videos [15]. The training optionally includes a set of responses corresponding to the samples and/or a mask of missing measurements. For classification, weight values might be given to the various classes of a dataset. Weight values given to each training sample can be used when improving the dataset based on accuracy feedback.

1.5 Spatiotemporal Data Extraction for Video Event Recognition

Automated video event recognition is amongst the most important goals of many video-based intelligent systems ranging from video surveillance to content-based video retrieval. It identifies and localizes video events characterized by spatiotemporal visual patterns of happenings over time, including object movements, trajectories, acceleration or deceleration, and behavior with respect to time constrains and logical flow. The semantics of video events that complement the features extracted by signal processing can be annotated using markup-based [16], ontology-based [17], and formal rule-based [18] representation.

The approaches to detect *regions of interest* (ROIs) fall into two categories. The *generalized approaches* are based on visual attention models that determine the likelihood of a human fixing his or her gaze on a particular position, such as the horizon, in a video sequence. Visual attention models are usually based on the features perceived by human vision, such as color, orientation, movement direction, and disparity, which can be combined into a saliency map to indicate the probability of pixel drawing attention [19]. Feature extraction can be extended with motion

detection and face detection to obtain more advanced visual attention models [20]. Since the positions at which the persons are looking are determined by the task [21], detection accuracy can be improved if the task is considered during detection. This is one of the main motivations behind the *application-based approaches*, which predict the region of interest a priori for a particular application (e.g., human faces in a video conferencing application).

Human action plays an important part in complex video event understanding, where an action is a sequence of movements generated by a person during the performance of a task. Action recognition approaches differ in terms of spatial and temporal decomposition, action segmentation and recognition from continuous video streams, and handling variations of camera viewpoint. Conventional 2D video scenes often do not convey enough information for *human action recognition* (HAR). When motion is perpendicular to the camera plane, the 3D structure of the scene is needed, which is typically obtained using depth sensors. The information obtained from depth sensors is suitable for human body model representation and skeleton tracking using masked joint trajectories through action template learning [22]. Human action recognition can also be performed using depth motion maps (DMMs), which are formed by projecting the depth frames of depth video sequences onto three orthogonal Cartesian planes and considering the difference between two consecutive projected maps under each projection view throughout the depth video sequence [23]. Real-time human action recognition is then utilized by a collaborative representation classifier with a distance-weighted Tikhonov matrix.

Beyond RDB-D depth cameras and Kinect depth sensors, inertial sensors are also applied in human action recognition [24]. The information fusion of the input obtained from RGB video cameras, depth cameras, and inertial sensors can exploit the benefits of the different representations of 3D action data [25].

In contrast to the static background of news videos, real-world actions often occur in crowded environments, where the motion of multiple objects distracts the segmentation of a particular action. One way to handle crowd flows is to consider them as collections of local spatiotemporal motion patterns in the scene, whose variation in space and time can be modeled with a set of statistical models [26].

1.6 Conceptualization of Multimedia Contents

While images and audio files might contain embedded machine-readable metadata, such as the geo-coordinates in JPEG image files [27] or the performer in MP3 audio files, video files seldom have anything beyond generic technical metadata, such as resolution and length, that do not convey information about the actual content. Because of the lack of content descriptors, the meaning of video scenes cannot be interpreted by automated software agents. Search engines still rely heavily on textual descriptors, tags, and labels for multimedia indexing and retrieval, because text-based data is still the most robust content form, which can be automatically processed using natural language processing.

The application of well-established image processing algorithms is limited in automated video processing, because videos are far more complex than images. Although there are common challenges in image and video processing, such as occlusion, background clutter, pose and lighting changes, and in-class variation, videos have unique characteristics. In contrast to images, videos convey spatiotemporal information, are inherently ambiguous, usually have an audio channel, are often huge in size, and are open to multiple interpretations. Moreover, the compression algorithm used in a video file determines which frames can be used, making a careful selection necessary to prevent processing frames depicting a transition (e.g., face washed out from motion blur). On top of these challenges, video surveillance applications and robotic vision require real-time processing, which is challenging due to the computing complexity and enormous amount of data involved.

One of the approaches to address some of the aforementioned issues is to use *Semantic Web standards* to create *ontologies*, which provide a formal conceptualization of the intended semantics of a knowledge domain or common sense human knowledge, i.e., an abstract, simplified view of the world represented for a particular purpose.

The logical formalism behind web ontologies provides robust modeling capabilities with formal model-theoretic semantics. These semantics are defined as a model theory representing an analogue or a part of the world being modeled, where objects of the world are modeled as elements of a set, and the relationships between the objects as sets of tuples. To accommodate the various needs of applications, different sets of mathematical constructors can be implemented for concrete usage scenarios while establishing the desired level of expressivity, manageability, and computational complexity. The formal foundation of ontologies provides precise definition of relationships between logical statements, which describes the intended behavior of ontology-based systems in a machine-readable form. The logical underpinning of web ontologies is useful not only for the formal definition of concepts, but also for maintaining ontology consistency and integrating multiple ontologies. This data description formalism does not make the *unique name assumption* (UNA), i.e., two concepts with different names might be considered equivalent in some inferred statements. In addition, any true statement is also known to be true, i.e., if a fact is not known, the negation of the fact cannot be implied automatically (*closed world assumption*, *CWA*).

Web ontologies hold sets of machine-interpretable statements, upon which logical consequences can be inferred automatically (as opposed to modeling languages such as the Unified Modeling Language (UML)), which enables complex high-level scene interpretation tasks, such as recognizing situations and temporal events rather than just objects [28]. The most common *inference*[1] tasks have been implemented in the form of efficient decision procedures that, depending on the

[1]Deriving logical conclusions from premises known or assumed to be true.

complexity of the formalism, can even guarantee that they will return a Boolean true or false value for the decision problem in a predetermined timeframe, i.e., they will not loop indefinitely.

Ontology engineers have covered many different knowledge domains for multimedia content, thereby enabling the efficient representation, indexing, retrieval, and processing of, among others, medical videos, surveillance videos, soccer videos, and tennis videos. Nevertheless, the semantic enrichment of multimedia resources, which provides efficient querying potential, often requires human cognition and knowledge, making automated annotation inaccurate or infeasible. Considering the millions of multimedia files available online, manual annotation is basically infeasible, even though several social media portals support *collaborative annotation*, through which annotations of multimedia resources are dynamically created and curated by multiple individuals [29]. Manual annotation has many drawbacks, clearly indicated by the misspelt, opinion-based, vague, polysemous or synonymous, and often inappropriately labeled categories on video sharing portals such as YouTube.

By implementing Semantic Web standards according to best practices, the depicted concepts, properties, and relationships can be described with high-level semantics, individually identified on the Internet and interlinked with millions of related concepts and resources in a machine-interpretable manner. To minimize the long web addresses in knowledge representation, the namespace mechanism is frequently used to abbreviate domains and directories, and reveal the meaning of tags and attributes by pointing to an external vocabulary that describes the concepts of the corresponding knowledge domain in a machine-processable format. For example, a movie ontology defines movie features and the relationship between these features in a machine-processable format, so that software agents can "understand" their meanings (e.g., title, director, running time) in a dataset or on a web page by pointing to the ontology file.

Interlinking the depicted concepts with related concept definitions puts the depicted concepts into context, improves concept detection accuracy, eliminates ambiguity, and refines semantic relationships. Among other benefits, organized data support a wide range of tasks and can be processed very efficiently.

1.7 Concept Mapping

Computational models are used to map the multimedia features to concept definitions, which can eventually be exploited by hypervideo applications, search engines, and intelligent applications. Semantic concept detection relies on multimedia data for training typically obtained though low-level feature extraction and feature-based model learning. The efficiency of concept detection is largely determined by the availability of multimedia training samples for the knowledge domain

most relevant to the depicted concepts. Because of the time-consuming nature of manual annotation, the number of labeled samples is often insufficient, which is partly addressed by semi-supervised learning algorithms, such as co-training [30].

While learning multimedia semantics can be formulated as supervised machine learning, not every machine learning algorithm is suitable due to the limited number of positive examples for each concept, incoherent negative examples, and the large share of overly generic examples. Oversampling can address some of these issues by replicating positive training examples, albeit it increases training data size and the time requirement of training. Another approach, undersampling, ignores some of the negative examples; however, this might result in losing some useful examples. To combine the benefits of the two approaches while minimizing their drawbacks, negative data can be first partitioned, then classifiers created based on positive examples and negative example groups [31].

The definition of semantics for the concepts depicted in a region of interest, keyframe, shot, video clip, or video, along with their properties and relationships, is provided by vocabularies and ontologies. Well-established common sense knowledge bases and ontologies that can be used for describing the concepts depicted in videos include *WordNet*[2] and *OpenCyc*.[3] General-purpose upper ontologies, such as *DOLCE*[4] and *SUMO*,[5] are also used in multimedia descriptions. Depending on the knowledge domain, correlations of concepts might be useful for improving the conceptualization of multimedia contents [32]. For example, a beach scene is far more likely to depict surfs, sand castles, and palm trees than a traffic scene, and so collections of concepts that often occur together provide additional information and context for scene interpretation. One of the most well-known ontologies to provide such predefined semantic relationships and co-occurrence patterns, although not without flaws, is the *Large-Scale Concept Ontology for Multimedia (LSCOM)* [33]. Class hierarchies of ontologies further improve scene interpretation. For example, the subclass-superclass relationships between concepts of an animal ontology make it machine-interpretable that a koala is a mammal, therefore both concepts are correct for the concept mapping of a depicted koala, only the first one is more specific than the second one. Moreover, multimedia concepts are usually not isolated, and multiple concepts can be associated with any given image or video clip, many of which are frequently correlated. For example, a koala is very likely to be depicted together with a eucalyptus tree, but more than unlikely with a space shuttle. In ontology-based scene interpretation, the a priori and asserted knowledge about a knowledge domain can be complemented by rule-based, inferred statements [34].

[2] http://wordnet-rdf.princeton.edu/ontology

[3] https://sourceforge.net/projects/texai/files/open-cyc-rdf/1.1/

[4] http://www.loa.istc.cnr.it/old/ontologies/DLP3971.zip

[5] http://www.adampease.org/OP/

1.8 Implementation Potential: From Search Engines to Hypervideo Applications

The machine-interpretable descriptions created using Semantic Web standards provide universal access to multimedia data for humans and computers alike. The unique identifiers used by these descriptions enable the separation of the description from the multimedia content, which is very beneficial in multimedia retrieval, because small text files are significantly easier to transfer and process than the actual multimedia files and encourage data reuse instead of duplicating data. The formal concept definitions eliminate ambiguity in these descriptors, and their interlinking makes them very efficient in finding related concepts. Furthermore, huge multimedia files have to be downloaded only if they seem to be truly relevant or interesting to the user. These descriptors can be distributed in powerful purpose-built databases and embedded directly in the website markup as lightweight annotations to reach the widest audience possible, providing data for state-of-the-art search engine optimization.

Semantics enable advanced applications that exploit formal knowledge representation. Such computer software can provide fully customized interfaces to service subscribers, automatically identify suspicious activity in surveillance videos, classify Hollywood movies, generate age rating for movies, and identify previously unknown risk factors for diseases from medical videos.

The semantically enriched multimedia contents can be searched using multimedia search terms, somewhat similar to searching text files. For example, users can find music that actually sounds similar (have similar frequencies, wavelengths, instruments, etc.) to the music they like. Videos can be searched for certain clips or a particular kind of movement. Hypervideo applications can play videos while displaying information about their content, position the playback to a particular part of a video based on semantics, and so on.

1.9 Summary

This chapter listed the main challenges of machine interpretation of images and video scenes. It highlighted the limitations of low-level feature-based classification, object recognition, and multimedia understanding. The utilization of the conceptualization of multimedia contents in search engines for content-based multimedia retrieval and hypervideo applications was also discussed.

http://www.springer.com/978-3-319-54065-8

Description Logics in Multimedia Reasoning

Sikos, L.F.

2017, XIII, 205 p. 25 illus., 20 illus. in color., Hardcover
ISBN: 978-3-319-54065-8