Exploitation of knowledge in video recordings

Alexia Briassouli and Ioannis Kompatsiaris

Multimedia Knowledge Lab Informatics and Telematics Institute,
6th km Charilaou-Thermi Road, 60361 Thessaloniki, Greece
{abria,ikom}@iti.gr,
http://mklab.iti.gr/

Abstract. Recently there has been great progress in hardware and communication technologies, which has created a large increase in the amount of multimedia information available to users. Multimedia applications become more useful as their content becomes more easily accessible, so new challenges are emerging in terms of storing, transmitting, personalizing, querying, indexing and retrieval of the multimedia content. Examples include the usage of multimedia data in business, entertainment, medicine, libraries, law and many other domains. For practical use, a description and deeper understanding of the information at the semantic level is required [1]. In this work, the exploitation of video processing and its combined use with knowledge is presented, for the extraction of a higher level understanding of the content.

1 Introduction

Initial attempts to extract higher level concepts, namely semantics, from multimedia data were based on manual textual annotation. However, these methods are extremely labor intensive, they suffer from subjectivity, and have inter-operability issues. For this reason, attention currently focuses on the automated, or semi-automated, extraction of semantics. An intermediate solution is to use automatic techniques that exploit textual information associated with multimedia content, when it exists. However, the multimedia data often contains additional information, which is not present in the textual content, so research has also followed the direction of processing the audiovisual data itself, in order to extract semantics. Moving from low-level perceptual features to high-level semantic descriptions that are relevant to human cognition, i.e. bridging the semantic gap, has formed what are known as content-based (analysis and) retrieval approaches, where focus is on extracting the most representative numerical descriptions and defining metrics that emulate the human notion of similarity. Low-level descriptors, metrics and segmentation tools are fundamental building blocks of any multimedia content manipulation technique, but they fail to fully capture the semantics of the audiovisual medium. For the successful analysis of multimedia content, low-level processing techniques are combined with a priori domain specific knowledge, leading to a high-level representation of multimedia content [2].

Depending on the adopted knowledge acquisition and representation process, two approaches can be identified in the relevant literature: implicit, realized by machine

learning methods, and explicit, realized using knowledge structures. Machine learning techniques have proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Moreover, these elegantly handle problems of high dimensionality. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) [3], [4]. On the other hand, knowledge-based approaches make use of prior knowledge in the form of explicitly defined facts and rules, i.e. they provide a coherent semantic domaina to support "visual" inference in the specified context [5], [6]. These facts and rules may connect semantic concepts with other concepts, or with low-level visual features.

In this work we initially present the capabilities of signal processing for the extraction of semantics from multimedia content (Sec. 2). The enrichment of this extracted information with a priori knowledge is presented in Sec. 3, while examples of applications of these techniques are provided in Sec. 4.

2 Multimedia Signal Processing

As stated in Sec. 1, multimedia signal processing can lead to the extraction of semantics from data in an implicit manner. The data is processed using signal processing algorithms, in order to extract characteristic discriminating features, which can lead to successful recognition and classification. The resulting recognition and classification algorithms can be applied to audio data in order to detect speakers, emotions, the locations of the source of speech. Visual data can be processed to detect and recognize objects, activities and humans, with very interesting and useful applications. Essentially, recognition systems acquire *implicit* knowledge by extracting features from multimedia input data, and classifying the observations, usually by comparing the input data to "training data" which has been previously classified. The features that are extracted can be related to audio processing data, to color, texture, motion in video, or a combination of them, depending on the application at hand.

One of the most commonly used schemes for classification are Support Vector Machines (SVMs), which are linear classifiers, that assign input data (of dimension d-1) to a hyperplane (dimension d). Another category of classification methods which are very popular are Hidden Markov Models (HMMs). HMMs model a process as being characterized by observable and hidden variables. They use training data in order to characterize the hidden variables according to a particular model. These model characteristics are then used for recognition of testing data, assuming that is of a similar nature to the training data. Fig. 1 shows some representative results of the recognition of concepts in images.

Video data can be processed to detect faces or humans, as well as objects or even concepts (such as "beach", "mountain" etc). Face detection algorithms locate human faces in multimedia data, and can be further employed in recognition systems [7]. Initial systems were limited to finding only frontal views of faces, but currently new methods are being developed that are able to detect faces rotated with respect to the viewer [8]. Human detection methods also exist, to localize entire humans in video [9], often based

on their appearance, e.g. information about their silhouette. When searching through multimedia content, reliable face or human detection algorithms can help find content with a particular actor. The extraction of a specific concept allows searching or grouping of data with that concept, for example videos containing beach scenes. Recognition and classification methods are developed for the detection, recognition and characterization of various objects (not only humans), and are also applied to the recognition of activities, and lately, of more complex events.

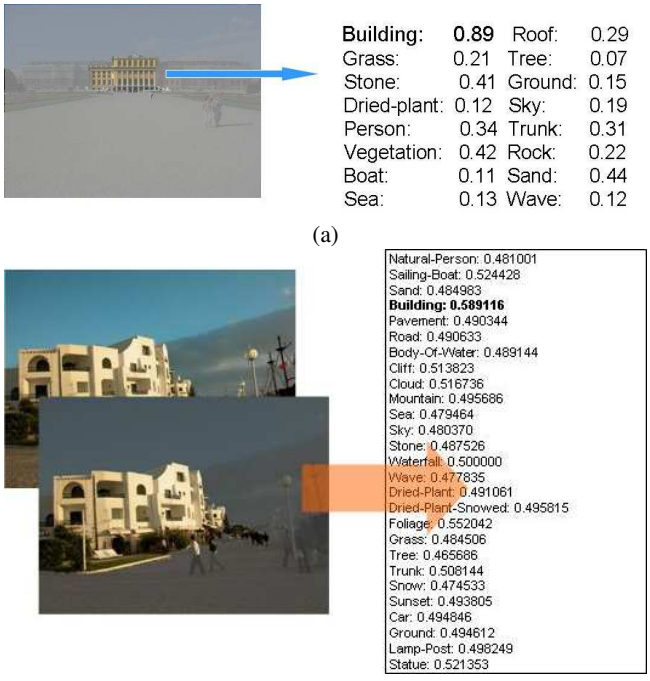

Fig. 1. Detection of concepts in images using SVMs.

The detection of activities and events in video data also makes use of machine learning for recognition. Since activities and events in video are characterized by motion, they are detected and classified after the extraction of motion characteristics from a video sequence, in addition to appearance features. Appearance features, such as the dominant color in a scene, have been used with success to separate videos into shots, to segment video frames and detect objects or humans in them. Motion features have been used as indicators of the amount of activity in a sequence, or to localize regions of activity [10]. The combination of motion and appearance features has led to the detection of particular events, for example in sports, or surveillance applications. Motion detection can also be extended to tracking, in order to find the locations of a moving

entity over a sequence of frames. This can be useful in traffic or surveillance applications, for example, where trajectories can be found, and "anomalous" behavior detected in the video.

Characteristic motion "signatures" may also be derived, providing implicit information about an event or activity taking place. For example, in Fig. 2, the processing of the motion in the video led to the extraction of a binary mask showing which pixels were active while the child threw the ball through the hoop and in Fig. 4 the characteristic motion of the tennis serve can be seen in the corresponding binary mask [11].

(a) (b)

Fig. 2. (a) Video frame of kid playing basketball. (b) Active pixels during the shot.

3 Knowledge

The introduction of explicit knowledge plays a significant role in bridging the semantic gap and extracting meaningful semantics from multimedia data. Among the possible domain knowledge representations, ontologies present a number of advantages. They provide a formal framework for supporting explicit, machine-processable semantics definitions, and they also enable the derivation of implicit knowledge through automated inference. Ontologies are a representation of a shared understanding about a domain and form an important part of the emerging Semantic Web since the latter is based on ontologies for enhancing (annotating) content with formal semantics. This will enable autonomic agents to reason about Web content and to carry out more intelligent tasks

on behalf of the user. Thus, ontologies are suitable for expressing multimedia content semantics so that annotation, automatic semantic analysis and further processing of the extracted semantic descriptions are allowed. Furthermore, ontologies provide a formal framework for exploiting the generated semantic descriptions in context representation, retrieval, personalization and related applications.

These advantages have recently led to the development of ontologies specifically for multimedia data. The complexity of multimedia data, in combination with the need for high-level semantic analysis, have turned the ontology-driven representation of information related to and concerning multimedia content into a rather demanding process.

For example, the main challenge in building a knowledge infrastructure for multimedia analysis and annotation is to link low-level multimedia properties, such as spatio-temporal multimedia document structure and semantic concepts in a clean, extensible, effective and efficient manner. A consensus is emerging that there is a need for the development of highly focused, easy to use, comprehensible, and non-overlapping multimedia ontologies. They should be kept simple and small, addressing the substantial needs of the applications/systems that are going to use them. Modularity, clarity and lack of ambiguities should characterize them. Similarly, requirements hold for other applications as well such as retrieval, personalization and context representation.

4 Overall System, Combined Approaches

Naturally, the optimal approach to extract semantics from multimedia data is to take advantage of all the knowledge available, i.e. both implicit, extracted from the data itself, and also explicit, which may be available a priori, e.g. in the form of ontologies. Such a system is depicted in Fig. 3.

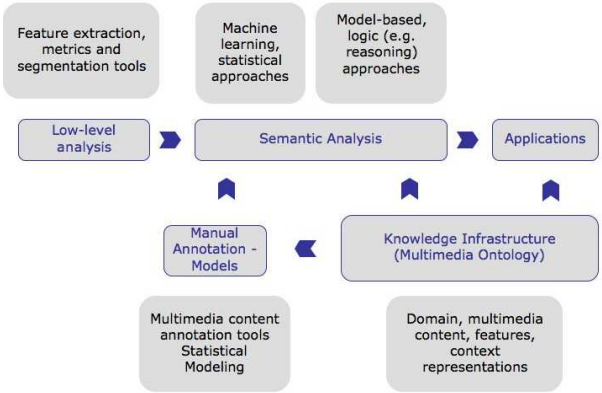

Fig. 3. Overall system combining multimedia analysis with knowledge.

Systems that combine explicit knowledge in the form of ontologies with implicit knowledge, extracted from multimedia data, have been developed for various applications. The sports domain has received much attention, as there are characteristic activities that take place in sports events, and also because there are specific rules characterizing each game. Fig. 4 shows an example of the extracted active pixels in a tennis game, with the characteristic shape of the player serving a shot.

(a) (b)

Fig. 4. (a) Video frame of tennis player serving. (b) Active pixels during the serve.

A system that combines the low-level features extracted from video data, in combination with information extracted from audio and text, could also be employed in a judicial setting, by processing data from court trials. In that case knowledge structures can be used to provide contextual information or to find how the concepts extracted at the low-level stage are connected to higher-level meanings. Few applications exist until now, that process multimodal data from video in order to extract semantics, although many signal processing algorithms are well suited for them. In court trials, people are present, so a face or human detection module [12], [13] is useful for localizing the people in the courtroom. Fig. 5 shows two examples of face detection in setups similar to those of an actual courtroom.

The motions taking place in a video of a courtroom can be also processed to extract useful semantics about what is taking place [14]. For example, a lawyer may be gesticulating more intensely when they are making a point. The movement of a witness, e.g. turning their head, may also indicate emotions such as fear, insecurity, which would be useful to be detected in a trial. Fig. 6 shows a person making characteristic gestures from which semantics can be inferred.

(a) (b)

Fig. 5. Face detection.

In order to take advantage of all the information available in the multimodal data, and infer the most meaningful semantics, a combined approach would be the most reasonable solution. This is because different modalities of the data often contain complementary information. For example, a video containing a person gesticulating contains information that is not present in the audio or transcript of that scene. Similarly, the emotions detected in the audio or visual recording of a person speaking, cannot be found in the text. In this section we shall present how the combined use of the multimodal processing can be of use in a judicial trial, where transcripts have been the main source of information until now. By processing video of the trial, emotions such as fear, anger or nervousness can be detected from minor gestures of the actors in the scene or from their facial expressions. Therefore, a component for face detection, that also provides information about facial expressions could provide information about emotions and enrich the transcript's contents. In parallel, the system should also contain a component which analyzes gestures, as well as more general kinds of motions. For example, the turning of a head is significant in a trial, while the detection of highly animated gestures indicates that an important part of the trial is taking place. This information may not be as easy to extract from the text or audio alone, and can help the users access only the interesting parts of the video. The video processing modules should, of course, be combined with audio and text components. The text analysis can immediately provide information about the actors in the trial, the part of the proceedings being analyzed and, of course, what is being said. The audio processing further enriches this information, as it may contain sounds that are not recorded in the transcript, and also be used to detect characteristic intonations, emotions etc. As explained in Sec. 3, the implicit knowledge each of the multimedia processing components extracts, is usually not at a high enough level to express more abstract concepts. For this reason, explicit knowledge would also be needed, to formulate a complete system. Explicit knowledge can be provided for

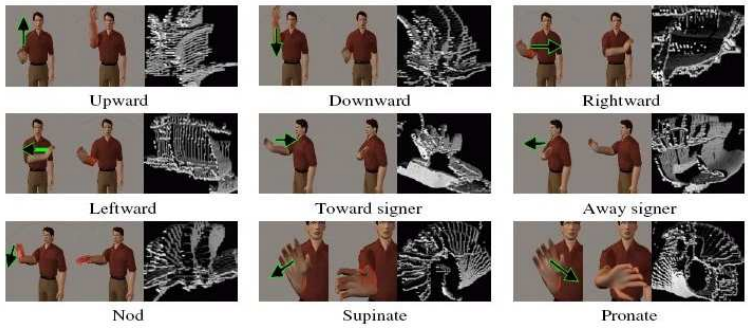

Fig. 6. Gesture recognition.

each modality, via ontologies tailored to the information derived from the audio, text and video separately. Finally, the resulting semantics can be combined in a unifying ontology, specifically designed for the needs of the application in question (a judicial trial in this example).

5 Conclusions

In this paper, we have demonstrated how modern multimedia processing systems have evolved to be able to extract useful semantics from their input data, be it video, audio, image and/or text. Low-level processing is useful as it leads to the extraction of characteristic features from the multi-modal data, that can enrich the information available, e.g. the transcript of a video. The knowledge derived by low-level processing is implicit, and fundamental for the full description of multimodal data. However, these results do not always correspond to unique, more abstract concepts, that are more commonly used by humans. For this reason, the incorporation of knowledge structures is particularly important in the construction of a complete system. The specific characteristics of an application can be expressed in such a knowledge structure, in order to arrive at the correct conclusions and semantics related to the data being processed. Such a system can be developed and used in a wide range of domains, including sports, surveillance, medicine, and judicial applications.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-214306 - JUMAS.

References

1. Chang, S.F.: The holy grail of content-based media analysis. IEEE Multimedia **9**(2) (2002)

2. Al-Khatib, W., Day, Y., Ghafoor, A., P.B., Berra: Semantic modeling and knowledge representation in multimedia databases. IEEE Transactions on Knowledge and Data Engineering **11**(1) (1999)
3. Assfalg, J., Berlini, M., Bimbo, A.D., Nunziat, W., Pala, P.: Soccer highlights detection and recognition using hmms. In: IEEE International Conference on Multimedia and Expo (ICME). (2005) 825–828
4. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ (2003)
5. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V., Strintzis, M.: Knowledge assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology (**15**(10)) 1210
6. Hollink, L., Little, S., Hunter, J.: Evaluating the application of semantic inferenc-ing rules to image annotation. In: 3rd International Conference on Knowledge Capture (K-CAP05). (2005)
7. Li, Z., Tang, X.: Bayesian face recognition using support vector machine and face clustering. In: CVPR04. (2004) II: 374–380
8. Huang, C., Ai, H., Li, Y., Lao, S.: High-performance rotation invariant multiview face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (**29**(4)) 671
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05). Volume 1. (2005) 886–893
10. Briassouli, A., Mezaris, V., Kompatsiaris, I.: Color aided motion-segmentation and object tracking for video sequences semantic analysis. International Journal of Imaging Systems and Technology (IJIST), Special Issue on Applied Color Image Processing (**17**(3)) 174
11. Briassouli, A., Mezaris, V., Kompatsiaris, I.: Color aided motion-segmentation and object tracking for video sequences semantic analysis. International Journal of Imaging Systems and Technology (IJIST), Special Issue on Applied Color Image Processing **17**(3) (2007) 174–189
12. Rowley, H., Baluja, S., Kanade, T.: Rotation invariant neural network-based face detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (1998)
13. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (**20**(1)) 23
14. Wong, S., Cipolla, R.: Continuous gesture recognition using a sparse bayesian classifier. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. (Volume 1.) 1084