Legal Multimedia Management and Semantic Annotation for Improved Search and Retrieval

Jorge González-Conejero, Emma Teodoro, Nuria Galera
Universitat Autònoma de Barcelona
Institute of Law and Technology
UAB-IDT, Cerdanyola del Vallès, 08290 SPAIN
{*jorge.gonzalez.conejero, emma.teodoro, nuria.galera*}*@uab.es*

Abstract. In this work, we study the possibilities of multimedia management and automatic annotation focused on legal domain. In this field, professionals are used to consume the most part of their time searching and retrieving legal information. For instance, in scenarios as e-discovery and e-learning search and retrieval of the multimedia contents are the basis of the whole applications. In addition, the legal multimedia explosion increases the need of store these files in a structured form to facilitate the access to this information in an efficient and effective way. Furthermore, the improvements achieved by sensors and video recorders in the last years increase the size of these files, producing an enormous demand of storage capability. JPEG2000 and MPEG-7 are international standards by the ISO/IEC organization that allow to reduce, in some degrees, the amount of data needed to store these files. These standards also permit to include the semantic annotation in the considered file formats, and to access to this information without the need to decompress the contained video or image. How to obtain the semantic information from multimedia is also studied as well as the different techniques to exploit and combine this information.

Keywords: Legal multimedia, semantic-based search and retrieval, JPEG2000 and MPEG-7

1. Introduction

Nowadays, legal professionals are used to consume an important part of their time searching, retrieving, and managing legal information. However, the recent explosion of multimedia legal contents has resulted in rising costs and requires more management capacities than ever before. Improving the functionalities for search, retrieval, and management of multimedia legal documents is paramount to fully unlock the potential of those contents for legal practice and to develop specific management solutions for different profiles of legal users (Brickell and Langer, 2009).

The multimedia files carries a meaning which can be very versatile. For a human the meaning of the message is immediate, but for a computer that is far from true. This discrepancy is commonly referred to as the *semantic gap* (Smeulders et al., 2000). Semantic multimedia annotation is the process of automatically detecting the presence of a concept in an image or video stream. In the literature, there are several works that address the multimedia annotation based on their meaning for different fields. In (Ballan et al., 2010)

an approach for automatic annotation and retrieval of video content is presented, based on ontologies, rule learning with first order logic, and semantic concept classifiers. An automatic video retrieval method based on high-level concept detectors is presented in (Snoek et al., 2007), defining a set of machine learned concept detectors enriched with semantic descriptions. In (Zha et al., 2007) a more general and comprehensive ontology to annotate video contents is described. Usually, an ontology consists of lexicon, properties, and relations. In this work LSCOM (Snoek et al., 2006) is used to construct the lexicon, describe concept property as the weights of different modalities which are obtained manually or by data-driven approach, and model two types of concept relations. The work (Gonzàlez et al., 2008) presents a Cognitive Vision System which explains the human behavior of monitored scenes using natural language texts. Here, the trajectories of human agents are obtained to generate textual interpretations of their motion, also inferring the conceptual relationship of each agent. The human behavior model is based on Situation Graph Trees.

Nevertheless, there is no available systems within the judicial domain to automatically index, tag, or annotate audiovisual files taking into account the requirements from judicial procedures. The annotation process for multimedia files produced by the judicial domain has several important benefits for law professionals. One of the most important features is that the annotation facilitates the search based on the meaning of the multimedia files, improving the legal frameworks and applications focused on, for instance, e-learning (Xin, 2009) and e-discovery (Baron and Thompson, 2007).

Figure 1 depicts an example of a general scheme to process multimedia files extracting semantic information. This scheme consists of three main stages. The first one is the extraction of events and concepts from videos and images. Exploiting high-level concepts and low-level descriptors we can achieve an automatic video retrieval method, see for instance (Snoek et al., 2007). The second stage applies a compression to the multimedia files to reduce the storage requirements and stores the semantic information in the same file, see (González-Conejero, 2010). The last stage is aimed to extract the main concepts of the user query to match the semantic information extracted from the multimedia files.

The aim of this work is to discuss the possibilities of annotate the legal multimedia contents and how these annotated files are managed. Recalling Figure 1, we study the composition of the three stages: annotation, management, and extracting concepts from the user query that fits into the legal multimedia field.

The paper is organized as follows: Section 2 explores the suitability of two international standards, JPEG2000 and MPEG-7, to manage multimedia files produced by the judicial procedures and the concepts and events extracted from these files; Section 3 studies how different features and semantic

events/concepts are extracted from the multimedia files; finally, Section 4 summarizes the work and points out some conclusions.

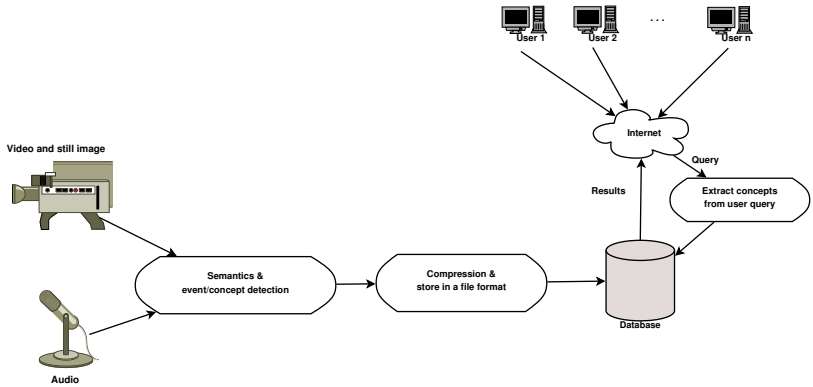

Figure 1. Main scheme for annotation of still image and video files.

2. Legal Multimedia Management

In this work, we are focused on the improving of search and retrieval applications in legal multimedia datasets. Nevertheless, how this multimedia files and the extracted semantic events/concepts are stored are also important topics. The widely use of multimedia files in the judicial domain and the improvements achieved by sensors and video recorders in the last years produce an enormous demand of storage capability. Images and videos usually contain highly redundant information, which can be exploited to compress and reduce, in some degrees, the amount of data needed to store these files. Apart from compression, the manipulation of multimedia currently requires other advanced features. Some of these features are the availability to transmit images and videos interactively over the network, to support error resilience or even to supply capabilities of watermarking and fingerprinting. Encoding systems must take these needs into account to provide a flexible framework that allows an efficient management.

A general description of the compression process is depicted in Figure 2. The input image is encoded, and the produced binary file is stored in a database (or similar) and/or transmitted over the network. Then, the original image is recovered at the client side through the decoder framework. In this Section, we describe two different coding systems that are able to: 1) achieve high coding performance; 2) provide an efficient management of the multimedia files; and 3) store semantic information in their specific file formats. JPEG2000 (Taubman and Marcellin, 2002) and MPEG-7 (Chang et al., 2001)

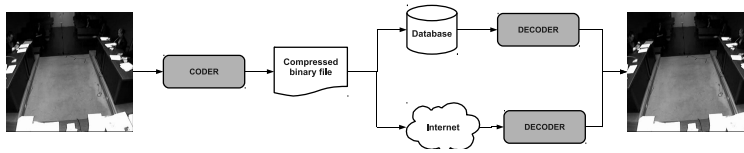

Figure 2. General compression scheme.

are standards from the International Standard Organization (ISO) and fulfills most of the requirements from the judicial field. The next sections contain a brief description of both coding systems.

2.1. JPEG2000

JPEG2000 is one of the latest standards developed by the Joint Photographic Experts Group (JPEG) and is structured in 13 different parts, addressing the encoding, transmission, security, and manipulation of still images and video. Since a description of the JPEG2000 is not the aim of this work, the interested reader is referred to: (Skodras et al., 2001) and (Rabbani and Joshi, 2002). Table I summarizes the most important parts of the JPEG2000 standard for the legal multimedia management. In our previous work, (González-Conejero, 2010), the suitability of these parts in the management of the legal multimedia contents and the inclusion of semantics in the JPEG2000 file formats are discussed. In addition, a centralized scheme to store in a database all of these files in a JPEG2000 file format is proposed.

Table I. Brief description of the 5 parts of the JPEG2000 standard suitable to manage legal multimedia contents.

– **Part 1 Core coding system**: description of the minimal decoder and a simple file format. It is the basis of the other parts.
– **Part 2 Extensions**: extensions of the core coding system, providing advanced coding features which can be used to enhance the coding performance or to manipulate unusual data types.
– **Part 3 Motion JPEG2000**: supports the manipulation of image sequences (motion).
– **Part 6 Compound image file format**: additional file format for tailored and compound documents.
– **Part 8 Secure JPEG2000**: description of a file syntax for interpreting secure image data and a normative process for registering security tools.
– **Part 9 Interactivity tools, APIs and protocols**: description of the transmission protocol JPIP, devised to interactively transmit JPEG2000 images.

2.2. MPEG-7

MPEG-7 is formally known as Multimedia Content Description Interface. While the prior standards focus on coding representation of audio and visual content, MPEG-7 focuses on description of multimedia content. MPEG-7 complements the existing MPEG standards suite and aims to be applicable to many existing formats, which include non-MPEG format and non-compressed formats as well. Table II summarizes the most important parts of the MPEG-7 standard suitable to manage legal multimedia files. In the literature there are several works that describe the different parts of the MPEG-7 standard, see for instance (Avaro and Salembier, 2001), (Hunter, 2001) and (Sikora, 2001).

Table II. Brief description of the 5 parts of the MPEG-7 standard suitable to manage legal multimedia contents.

– **Part 1 Systems**: specifies system level functionalities, such as preparation of MPEG-7 descriptions for transport/storage, synchronization of content descriptions, and development of conformance decoders.
– **Part 2 Description Definition Language**: is a derived by extension of XML schema to address other requirements specific to MPEG-7.
– **Part 3 Visual**: specifies features such as color, texture, shape and motion. Other elements required are structure, viewpoint, localization, and temporal.
– **Part 4 Audio**: addresses different classes of audio.
– **Part 5 Multimedia Description Schemes**: specifies a high-level framework that allows generic description of all kinds of multimedia.

3. Multimedia annotation

The need to review documents imposes considerable overhead in terms of cost and time, and challenges the capacity for legal system to perform search and retrieval matters effectively. In Spain, the Civil Procedure Act of January 7th, 2000 (1/2000) introduces the video recording of oral hearings. Consequently, Spanish civil courts are currently producing a massive number of multimedia files that have substituted the written transcripts and have become part of the judicial file, together with suits, indictments, injunctions, judgments and pieces of evidence. Lawyers, prosecutors and judges need to access these contents when preparing similar cases or when appealing to superior courts.

Inclusion of semantic fields and the semantic-based search and retrieval has been one of the long-term goals of multimedia computing to bridge the above mentioned *semantic gap*. In addition, the automatization of this annotation is an important feature in the legal multimedia domain due to the enormous quantity of multimedia files generated. Nowadays, this annotation has to been done manually, and the excess of work of the different employees in the court could penalize this step. Next sections studies the state-of-the-art of the multimedia annotation files and the suitability for the legal multimedia domain. Here, how the files are automatically annotated are described. Another important topic is how to match the query of the user and the annotation of the multimedia files (semantic gap from user query).

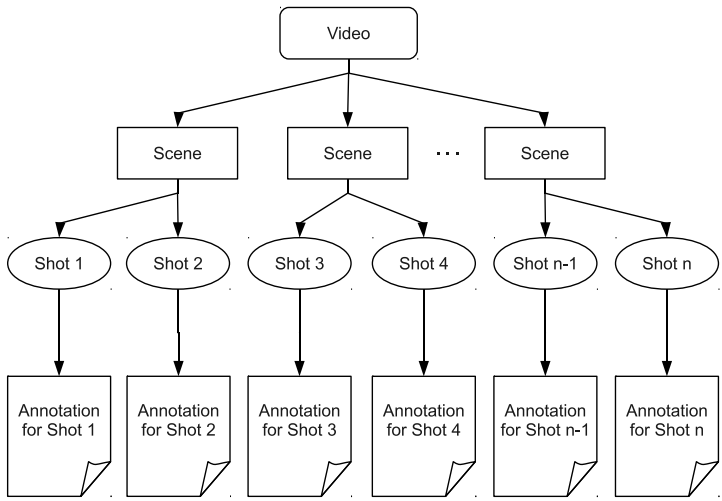

Figure 3. The video is divided in different shots concerning the meaning of every scene.

3.1. VIDEO ANNOTATION

The video annotation begins with the detection of events and concepts. Here, video events are defined as the interesting events which capture the users' attention (i.e. something that happen in the video as a "car accident" or "goal in a soccer match"); whereas the concepts refer to high-level semantic features, like "news", "sports", etc. Most of the state-of-the-art event detection frameworks were conduced toward the videos with poor structure or without story units, such as surveillance and medical videos (Zhu et al., 2005). On the other hand, the concept extraction scheme is widely used on videos which have structured contents, for instance broadcast news. Another example of field

that has structured contents is the judicial domain, where its procedures are divided into ruled different parts that compose the entire process. However, the routines and constrains of daily practice in court result in a far more complex typology, revealing interesting differences between the formal provisions of the law and the actual development of the procedures. The *e-Sentencias* project (Casanovas et al., 2009) made an important effort to tackle this issue, defining an scheme of the typology of civil hearings in Spain as emerged in daily practice. The e-Sentencias also presents a framework to annotate and facilitates the navigation of the user across the different recordings of judicial oral hearings.

Semantic annotation involves temporal partitioning of the video sequence into meaningful units which serve as the basis for concept extraction and semantic annotation. Every meaningful part of the video is named as *shot* and it will be the minimum self-contained, well-defined and accessible unit. There are several works in the literature that address this issue, for instance (Amiri and Fathy, 2009) and (Meng et al., 2009). Shots are annotated with the semantic concept that each scene represents in their space of time. Figure 3 depicts a scheme of this process.

Approaches for deriving semantics based on low level features, such as color, texture and local descriptors, have shown their limitations in bridge the semantic gap. Modern approaches enable semantic search by generating a set of concept detectors to extract semantics from low level features. In (Hauptmann et al., 2007) how many concepts would be needed, and how they should be selected and used is studied. For different simulations in a broadcast news dataset they find that good retrieval can be achieved even when detection accuracy is low, if sufficiently many concepts are combined. Whereas the low level concepts are determined by the video features, concepts and detectors that learn from the mapping between a set of low level visual features and concept from examples have to be designed. The common idea in this topic is to apply a machine-learning technique, usually a Support Vector Machine (SVM), to automatically learn this mapping from the data. Other popular solutions is to apply a Bag-of-Words approach (Sivic and Zisserman, 2003) in which an image or video frame is represented as a bag of quantized descriptors referred to as visual-words. Then, this representation is used to compute histograms of visual-words frequencies used to train appropriate classifiers. Another approach proved for detection of specific object classes as “face” or “person” are (Viola and Jones, 2004) and (Jamieson et al., 2010).

On the other hand, exploitation of the semantic relationships between concepts is receiving a large attention from the researchers in this field, due to it can improve the detection accuracy of concepts and provides a richer semantic annotation of a video. Ontologies are expected to improve the computer systems detection even complex concepts and events from visual data. They organize semantic heterogeneity of information, using a formal representa-

tion, and provide a common vocabulary that encodes semantics and supports reasoning. Several works in the literature add ontologies combined with other features to improve the concepts detection. For instance, (Zha et al., 2007), (Wei et al., 2008) and (Snoek et al., 2007).

Figure 4 depicts a scheme of the hierarchy of the semantic annotation. From bottom to top of the figure: first step contains the multimedia files, from these files we can extract low-level descriptors as pixels, textures, speech recognition, etc. In the upper level, concepts used to the semantic annotation are defined. Finally, the user level contains the user query. Another important issue in semantic search and retrieval field is how to manage the user query. The accuracy of the user describing the query text to match the concepts extracted is a well-known problem. The richness of the vocabulary is also a problem for humans describing video in words. A variety of terms are used to describe the same video fragment by different users, or by same user in different contexts. Use ontologies to structure terms employed by users can make descriptions more consistent and can aid the user in selecting the terms for a semantic search.

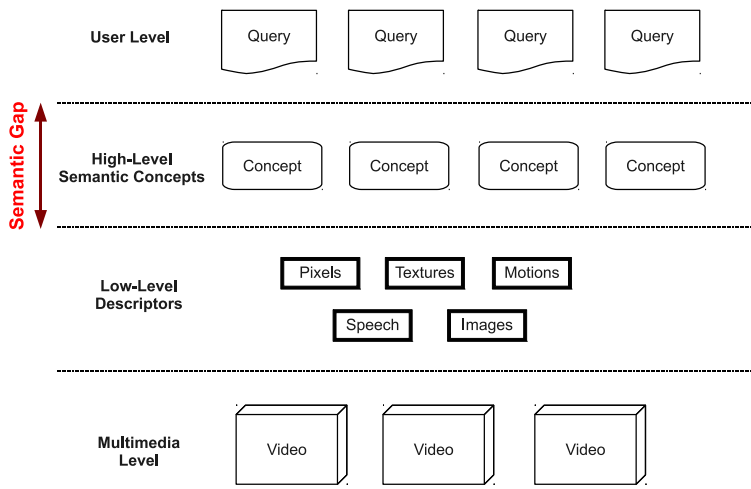

Figure 4. Video annotation scheme.

3.2. Still image annotation

The annotation of still images is a more complex process than the video annotation. The semantic meaning is more easy to extract when a complete scene could be analyzed. Nevertheless, in a still image the only information is the objects within the image. Image annotation can not be performed by simply

manually associating words to each image, firstly because it would be a very tedious task with the exponential increasing quantity of digital images and secondly because their content can not be fully described by a list of words.

Extraction of visual information directly from the images is required, nevertheless, bridging the semantic gap between the target semantic classes and the available low level visual descriptors is an unsolved problem. Consequently, it is crucial to select an appropriate set of visual descriptors that capture the particular properties of a specific domain and the features of each image class. For instance, local color descriptors, global color histograms, edge direction histograms, etc. The second crucial problem is to combine the low level descriptors in such a way that the results obtained with individual descriptors are improved.

In the literature, there are systems designed to learn meaningful correspondences between words and appearance models from cluttered images of multiple objects. Many approaches associate a caption word with a probability distribution over a feature space dominated by color and texture. This type of representation is less reliant on perceptual grouping than a shape model or a structured appearance model due to color and texture are robust to segmentation errors and the features configuration is not critical. There are several works that address the learning of configurations and problems of perceptual grouping. In (Barnard et al., 2003) a ranking scheme for potential merges of regions based on a similarity of word-regions associations. In a similar fashion, (Quattoni et al., 2007) use the co-occurrence of caption words and visual features to merge together equal features. Nevertheless, these models contain no spatial relationships between parts that would allow them to represent true part configurations. The work (Carbonetto et al., 2004) can successfully recognize a set of adjacent regions with widely varying appearance as being associated with a given word. The multiresolution statistical model introduced in (Li and Wang, 2003) can represent configurations of visual features across multiple scales. Here, each semantic class is associated with a layout for the entire image, where the division in parts is predefined. However, this system does not perform grouping. Other works avoid the perceptual grouping problem by focusing on domains where exists detailed prior knowledge of the appearance of the objects of interest, as in the task of matching names with faces, for instance see (Viola and Jones, 2004).

3.3. Multimedia Annotation and MPEG-7

All the information generated by the annotation process have to be stored in the multimedia files under a concrete file format, as we stated in Section 2. Figure 5 depicts a brief scheme of how the semantics/metadata is stored in concrete parts of the final file format. In the literature, most of the works concerning this issue pose the problem of the inclusion of semantics/metadata

in the MPEG-7. So, in this section we are focused on this standard that store the information produced by the annotation process carried out to the multimedia files. MPEG-7 can be used to create complex and comprehensive metadata descriptions of multimedia content. It is also defined in terms of an XML schema, however, the semantics have no formal grounding. There are description tools for diverse types of annotations on different semantic levels, ranging from very low-level features, such as visual or audio, to more abstract descriptions. The flexibility of MPEG-7 relies on structuring tools, which allow descriptions to be associated with arbitrary multimedia segments or regions, using different levels of abstraction.

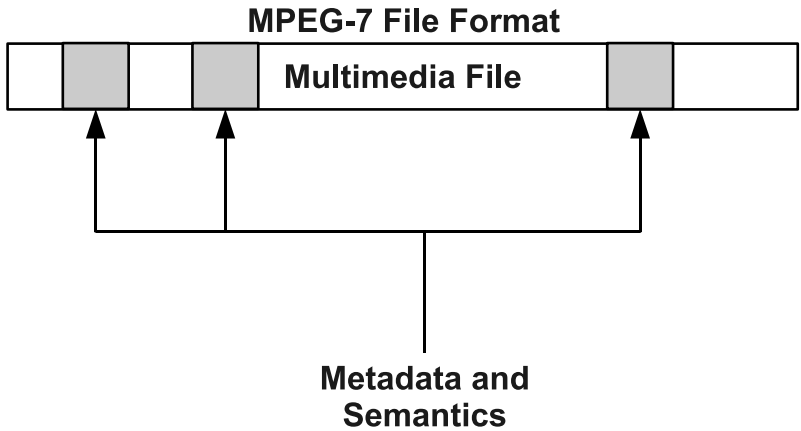

Figure 5. The semantic information is stored in the MPEG-7 file format.

Several works that relate the MPEG-7 standard and the semantics are presented in the literature. (Graves and Lalmas, 2002) proposes a model for video retrieval based upon the inference network model. The document network is constructed using video metadata encoded through MPEG-7 and captures information pertaining to the structural, conceptual and contextual aspects. For image classification, in (Spyrou et al., 2005) three content-based techniques based on fusing various low-level MPEG-7 visual descriptors are presented. One of this three techniques is based on neurofuzzy network, in this case fuzzy rules can be extracted in an effort to bridge the semantic gap between the low-level descriptors and the high level semantics of the image. In the video databases field, videos have to be presented in a compact and discriminative way to perform an efficient matching and retrieval of documents. In (Bertini et al., 2006) a method to obtain video representation to pose this issue is presented and it is based on features and descriptors taken from the MPEG-7 standard. Finally, (Bailer et al., 2006) proposes an approach for expressing semantics explicitly by formalizing the semantic constraints of a

profile using ontologies and rules, enabling interoperability and automatic use for MPEG-7 based applications.

4. Summary

Legal professionals consume the most part of their time searching and retrieving legal information. Furthermore, the explosion of legal multimedia contents in the judicial domain produces an enormous quantity of this files that have to been stored in a way that facilitates the search process. E-discovery and e-learning are fields that also need to store the information in a structured manner to improve the search and retrieving applications. This work is aimed to study the automatic annotation of legal multimedia contents based on their semantic meaning, and how this annotation is stored in an international standard file format as JPEG2000 or MPEG-7. Taking advantage of both international standards, a compression process is also applied to reduce the amount of information needed to store all the multimedia files.

The JPEG2000 coding system is one of the latest standards proposed by the Joint Photographic Experts Group (JPEG). It is composed by 13 different parts addressing the encoding, transmission and manipulation of still images and video. Six different parts of the JPEG2000 are suitable to manage legal multimedia files. On the other hand, the Moving Pictures Experts Group (MPEG) committee has developed many standards for still image and video compression. The last one is the MPEG-7 that improves the MPEG-4 including: Description Definition Language, audio-visual Descriptor, and Description Schemes. All of them aimed to define, at different levels, syntax and semantic capabilities. While the prior standards from the MPEG committee focus on coding audio and visual content, MPEG-7 focuses on multimedia.

Semantic-based search and retrieval has been an important issue for years in the multimedia computing universe. Multimedia files have associated a meaning which can be very versatile. Semantic multimedia annotation is the process of automatically detecting the presence of a concept in an image or video stream. The annotation of video and images have similar problems, however, to annotate an image is more complex than a video due to videos have scenes that facilitates the detection of high-level concepts and events. Another important topic of research in this field is the extraction of concepts from the user query, where ontologies can help users to describe videos in words.

The annotation process of a video begins with a segmentation that split the video in different temporal partitions, named as "*shots*", in function of the different meaning of scenes in the video. Most works in the literature try to bridge the semantic gap between the low-level descriptors (such as texture, pixel, audio, etc.) and high-level concepts (such as sport, news, car accident,

etc.). Although the use of a machine-learning technique as a Support Vector Machine (SVM) provides good results, other approaches, for instance either Bag-of-Words, or detection of different classes (faces, outdoor/indoor, etc) are used too, providing them acceptable results. In addition, the exploitation of semantic relationships between concepts is also an important field, since it can improves the detection accuracy. In the annotation of still images there are two crucial problems. The first is to select an appropriate set of visual descriptors that capture the particular properties of a specific domain and the features of each image class. The second crucial problem is how to combine the low-level descriptors in a fashion that the results obtained with individual descriptors are improved. There are several approaches to pose this problem such as models based on either the shape or structured appearance, or a distribution probability associated with a caption word. Other works address the learning of configurations and problems of perceptual grouping.

Finally, we can conclude that the management of legal multimedia files through the standards JPEG2000 and MPEG-7 can improve the storage requirements and the semantics/metadata management produced by the annotation process. Through both standards, the amount of storage requirements is reduced and the semantics/metadata information added to the final file could be accessed without the need of decompress the whole image or video. Furthermore, the existent techniques to carry out the semantic-based annotation of legal multimedia files can improve the search and retrieval applications for different legal fields as e-discovery and e-learning. The state-of-the-art summarized in this work is fully applicable to the legal domain. However, the particularities of the judicial procedures, for instance, the closed structure of oral hearings, have to been taken into account to improve the concept detection.

Acknowledgements

This work has been partially supported by the Spanish Government (Ministerio de Ciencia e Innovación and Ministerio de Industria, Turismo y Comercio) under Grants TSI-020501-2008-131 and CSO2008-05536.

References

Brickell, J. L. and Langer, A. M. (2009). Adapting to the Data Explosion: Ensuring Justice for all. In *IEEE International Conference on Systems, Man and Cibernetics Society*, pages 86–90. IEEE.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Ballan, L., Bertini, M., Del Bimbo, A., and Serra, G. (2010). Video Annotation and Retrieval Using Ontologies and Rule Learning. *IEEE Multimedia*, In press.

Snoek, C. G., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., and Worring, M. (2007). Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986.

Zha, Z.-J., Mei, T., Wang, Z., and Hua, X.-S. (2007). Building a comprehensive ontology to refine video concept detection. In *Proceedings of the international workshop on multimedia information retrieval*, pages 227–236.

Snoek, C. G., Worring, M., van Gemert, J., Geusebroek, J., and Smeulders, A. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA'06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, NY, USA.

Gonzàlez, J., Rowe, D., Varona, J., and Roca, F. X. (2008). Understanding dynamic scenes based on human sequence evaluation. *Image and Vision Computing*, 27(10):1433 – 1444.

Xin, C. (2009). E-learning Applications and Challenges. In *International Conference on Future Information Technology and Management Engineering*, pages 580–583. IEEE.

Baron, J. R. and Thompson, P. (2007). The Search Problem Posed by Large Heterogeneous Data Sets in Litigation: Possible Future Approaches to Research. In *International Conference on Artificial Intelligence and Law*, pages 141–147. ACM.

González-Conejero, J. (2010). Legal Multimedia Management through JPEG2000 framework. *Lecture Notes in Artificial Intelligence*, AI approaches to the complexity of legal systems(In press).

Taubman, D. and Marcellin, M. (2002). *JPEG2000: Image Compression Fundamentals, Standards, and Practice*, volume 642. Kluwer International Series in Engineering and Computer Science.

Chang, S.-F., Sikora, T., and Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695.

Skodras, A., Christopoulos, C., and Ebrahimi, T. (2001). The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58.

Rabbani, M. and Joshi, R. (2002). An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication*, 17(1):3–48.

Avaro, O. and Salembier, P. (2001). Mpeg-7 Systems: Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):760–764.

Hunter, J. (2001). An Overview of the MPEG-7 Description Definition Language (DDL). *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):765–772.

Sikora, T. (2001). The MPEG-7 Visual Standard for Content Description – An Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):696–702.

Zhu, X., Wu, X., Elmagarmid, A., Feng, Z., and Wu, L. (2005). Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):665–677.

Casanovas, P., Binefa, X., Gracia, C., Teodoro, E., Galera, N., Blázquez, M., Poblet, M., Carrabina, J., Monton, M., Montero, C., Serrano, J., and López-Cobo, J. M. (2009). *Law, ontologies and the semantic web: Channeling the Legal Information Flood*, volume 188 of *Frontiers in Artificial Intelligence and Applications*, chapter The e-Sentencias Prototype: A Procedural Ontology for Legal Multimedia Applications, pages 199–219. IOS Press, Amsterdam, Netherlands.

Amiri, A. and Fathy, M. (2009). Video Shot Boundary Detection Using QR-Decomposition and Gaussian Transition Detection. *EURASIP Journal on Advances in Signal Processing*, 2009(Article ID 509438):1–12.

Meng, Y., Wang, L.-G., and Mao, L.-Z. (2009). A shot boundary detection algorithm based on Particle Swarm Optimization Classifier. In *International Conference on Machine Learning and Cybernetics*, volume 3, pages 1671 – 1676, Baoding, China.

Hauptmann, A., Yan, R., Lin, W.-H., Christel, M., and Wactlar, H. (2007). Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News. *IEEE Transactions on Multimedia*, 9(5):958–966.

Sivic, J. and Zisserman, A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the Ninth International Conference on Computer Vision*, volume 2, pages 1470–1477, Nice, France. IEEE Computer Society.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Jamieson, M., Fazly, A., Stevenson, S., Dickinson, S., and Wachsmuth, S. (2010). Using Language to Learn Structured Appearance Models for Image Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):148–164.

Wei, X.-Y., Ngo, C.-W., and Jiang, Y.-G. (2008). Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces. *IEEE Transactions on Multimedia*, 10(6):1085–1096.

Barnard, K., Duygulu, P., Guru, R., Gabbur, P., and Forsyth, D. (2003). The Effects of Segmentation and Feature Choice in a Translation Model of Object Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–682.

Quattoni, A., Collins, M., and Darrell, T. (2007). Learning visual representations using images with captions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE CS.

Carbonetto, P., de Freitas, N., and Barnard, K. (2004). A Statistical Model for General Contextual Object Recognition. In *European Conference Computer Vision*.

Li, J. and Wang, J. (2003). Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(9):1075–1088.

Graves, A. and Lalmas, M. (2002). Video Retrieval using an MPEG-7 Based Inference Network. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346.

Spyrou, E., Le Borgne, H., Mailis, T., Cooke, E., Avrithis, Y., and O'Connor, N. (2005). Fusing mpeg-7 visual descriptors for image classification. *Lecture Notes in Computer Science*, 3697:847–852.

Bertini, M., Del Bimbo, A., and Nunziati, W. (2006). Video Clip Matching Using MPEG-7 Descriptors and Edit Distance. *Lecture Notes in Computer Science*, 4071:133–142.

Bailer, W., Hausenblas, M., Hofmair, P., and Schalatte, R. (2006). Enabling multimedia metadata interoperability by defining formal semantics of mpeg-7 profiles. *Lecture Notes in Computer Science*, 4306:41–55.

Author Index