MPEG-7 audio-visual indexing test-bed for video retrieval

L. Gagnon[*a], S. Foucher[a], V. Gouaillier[a], C. Brun[a], J. Brousseau[a], G. Boulianne[a], F. Osterrath[a], C. Chapdelaine[a] , J. Dutrisac[b], F. St-Onge[b], B. Champagne[c], X. Lu[c]

[a]R&D Department, Computer Research Institute of Montreal (CRIM),
550 Sherbrooke West, Suite 100, Montreal (QC), Canada, H3A 1B9

[b]Communications & Outreach Development Department,
National Film Board of Canada (NFB-ONF),
P. B. 6100, Centre-ville Station, Montreal (QC), Canada, H3C 3H5

[c]Department of Electrical Engineering, McGill University,
3480 University Street, Montreal (QC), Canada, H3A 2A7

ABSTRACT

This paper reports on the development status of a Multimedia Asset Management (MAM) test-bed for content-based indexing and retrieval of audio-visual documents within the MPEG-7 standard. The project, called "MPEG-7 Audio-Visual Document Indexing System" (MADIS), specifically targets the indexing and retrieval of video shots and key frames from documentary film archives, based on audio-visual content like face recognition, motion activity, speech recognition and semantic clustering. The MPEG-7/XML encoding of the film database is done off-line. The description decomposition is based on a temporal decomposition into visual segments (shots), key frames and audio/speech sub-segments. The visible outcome will be a web site that allows video retrieval using a proprietary XQuery-based search engine and accessible to members at the Canadian National Film Board (NFB) Cineroute site. For example, end-user will be able to ask to point on movie shots in the database that have been produced in a specific year, that contain the face of a specific actor who tells a specific word and in which there is no motion activity. Video streaming is performed over the high bandwidth CA*net network deployed by CANARIE, a public Canadian Internet development organization.

Keywords: MPEG-7, multimedia asset management, audiovisual indexing, content-based image retrieval, speech recognition, signal segmentation

1. INTRODUCTION

The purpose of this paper is to report about the development status of a modular and scalable MPEG-7 Audio-visual Documentation Indexing System (MADIS) with search capacities packaged into a demonstrable test-bed, for research and development on content-based retrieval of films. The outcome of the project will be a Web site integrated within the Canadian National Film Board (NFB) Cineroute site. NFB's mission is to produce and distribute distinctive, culturally diverse, challenging and relevant audiovisual works that provide Canada and the world a unique Canadian perspective. The organization has more than 10 000 archived audiovisual documents and 4 000 hours of stock shots, from 1895 to now. One hundreds (100) hours of new material is added each year. Manual logging and indexing is done since 1978 according to film subject, genre and related terms (thesaurus). The MADIS demo will run over CA*net, a national Internet research and education network deployed by the Canadian Internet development organization CANARIE Inc

Content-based search of audio-visual documents in large databases requires content-based indexing. This cannot be done only manually. Some sort of machine assistance is necessary in order to efficiently index and retrieve relevant video materials. Adapting a recent quote [1], one could say that indexing audio-visual documents relates to three basic

[*] lgagnon@crim.ca; phone 514 840-1235; fax 514 840-1244; crim.ca

issues: (1) which audio-visual features to index for a given application (e.g. the names of the actors, the spoken words and the scene type for video footage), (2) how to extract them (e.g. neural network classifier for face recognition, spectral features for speech recognition, etc.) and (3) how to organize the index table. Most of the scientific works done so far address the second issue, and most of the time for one modality, either visual, auditory or textual (see, for instance, [1-4] and references there-in). Combined multi-modal or at least audio-visual approaches have been addressed only recently (see for instance, [1,2] and [5,7] and references there in) with few commercial exploitations [8]. On the other hand, practical applications of content-based audio-visual indexing/retrieval necessitate taking into account the two other issues. Identifying appropriate features to be encoded relates to a careful analysis of the use cases with the end users. Selecting the indexing organization relates more to a content-based descriptor scheme like the one proposed in the MPEG-7 standard [9].

MADIS is designed to (1) be fully MPEG-7 compliant, (2) address both encoding and search, (3) combine audio, speech and visual modalities and (4) have search capability on the Internet. Among the 43 international project compiled by the MPEG consortium, only few address all those specifications at the same time [10,11]. MADIS offers the opportunity to address "hidden" scientific/technical issues that are very difficult to predict in such generic application. In order to minimize the technical risk, we target to a specific end-user application, that is, documentary films. It is thus an application-oriented project, for which 1) use cases are closely analyzed before the interface design, and 2) audiovisual feature extraction tools are selected and adapted from CRIM heritage [12, 13] or state-of-the-art open source codes.

In the current test-bed, the film encoding is done off-line semi-automatically. A XML file, compatible with MPEG-7/XML schema, is created for each film. In addition to the content management information, each XML file contains tags related to the audio and visual contents identified after the use cases analysis. The description decomposition is based on a temporal decomposition into visual segments (shots), key frames and audio/speech sub-segments. For each shot, key frames are automatically identified [14] and semantically clustered to allow frame retrieval by image category. Specific visual and audio characteristics are extracted and encoded for each shot segment. Visual features include global texture and color content, face detection and motion activity. Face detection and recognition are based on boosted cascade of simple classifiers [15, 16] and Hidden Markov Model (HMM) [17]. The audio band within a visual shot is segmented in speech and non-speech segments. An automatic speech recognition system developed at CRIM [13] is used to encode the spoken words in phone lattices. Non-speech segments are further segmented into three classes: music, silence and other sounds.

Audio-visual requests are done through a Java GUI with simple and advanced XQuery-based search engine capabilities. Management, visual and audio content are done through key words or images and combined according to the use cases specifications. For instance, an end-user can ask to point on movie shots in the database that have been produced in a specific year, that contain the face of a specific actor who tells a specific word and in which there is no motion activity. Video streaming is done over the high bandwidth CA*net network. Test-bed performance will be evaluated for audiovisual encoding accuracy, search time in XML trees and usability.

The paper is organized as follows. Section 2 recalls the main objectives and characteristics of the MPEG-7 standard in the context of audio-visual archiving. Section 3 presents the overall system architecture of the MADIS project with a special attention to the encoding, query and data flow aspects. Section 4 describes in more details the core technical encoding and retrieval modules. System usage is demonstrated in Section 5, followed by a conclusion.

2. MPEG-7

MPEG-7 (ISO/IEC International Standard 15938) was developed by the Moving Picture Experts Group. Formally named “Multimedia Content Description Interface”, MPEG-7 standard defines a normative indexing of multimedia content at many level ranging from low-level features to higher semantic description [9]. It allows to record information on both content management (media description, navigation and access, user interaction) and content itself (structure and semantics). Only the description structure is part of the standard; production or use of MPEG-7 descriptions is not.

MPEG-7 can address different media types in various formats and offer a generic framework to support a broad range of applications that necessitate the interpretation of multimedia content (e.g. content-bases retrieval, content management, navigation, filtering, and automated processing). It enables interoperability between applications and systems, which processes and manage multimedia content. The benefits of its use are numerous. MPEG-7 is harmonised with other standards that have demonstrated success and acceptance in both traditional media and new media businesses like W3C (XML, XML Schema), IETF (URI, URN, URL), Dublin Core, SMPTE Metadata Dictionary, etc. [11].

An MPEG-7 description is an XML file instantiating a subset of predefined normative tools. These tools are of two types: Descriptors (D), that define the XML syntax and the semantics of each feature of the multimedia content, and Description Schemes (DS), that assemble D and DS by specifying the structure and semantics relationships between them. These tools are defined by the Description Definition Language (DDL), which is based on the XML Schema and that allows creation of new DS or extension of existing ones. A set of systems tools is also available to support binarisation, multiplexing, synchronisation, transport and storage of descriptions as well as the management and protection of intellectual property.

MPEG-7 comprises 25 tools specific to the description of visual content including still images, video and 3D models. These tools can represent visual attributes such as colour, texture, shape, motion, and face features. The collaborative development of the standard has produced the eXperimentation Model (XM) software, which is a simulation platform for the MPEG-7 tools [18]. Its purpose is to provide a common framework to test MPEG-7 applications. Besides the normative components, XM also implements non-normative components as feature extraction and search capabilities. To each D or DS corresponds a set of applications divided in two types: the server (extraction) applications and the client (search, filtering) applications. Only the output produced by XM is normative. The methods implemented in XM are not part of the standard and can be adapted/changed as needed.

3. ARCHITECTURE

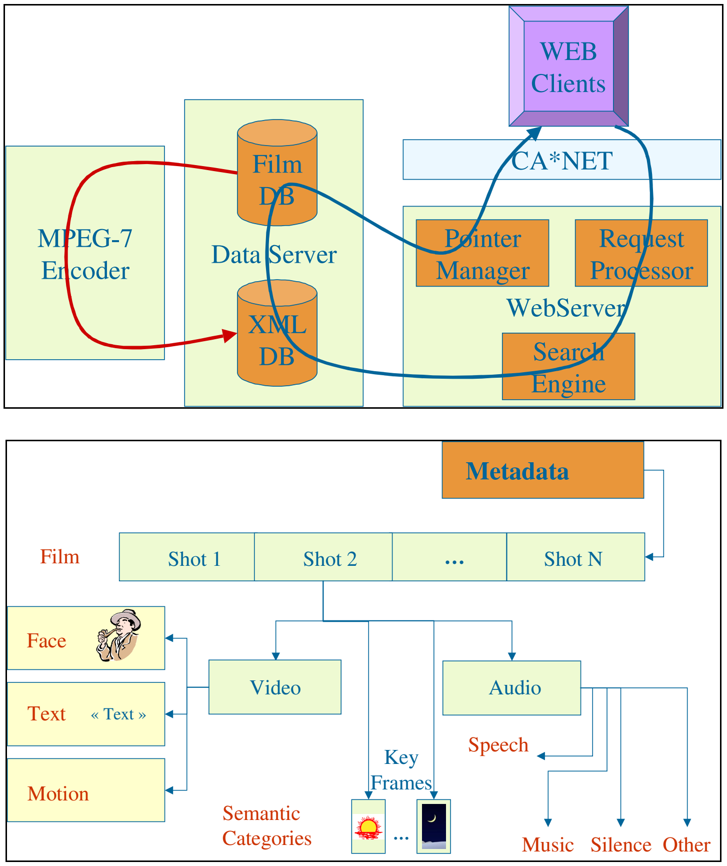


Figure 1: High-level block diagram showing the architecture of MADIS (top)
Description structure used in MADIS (bottom)

Figure 1 shows a high-level block diagram of MADIS architecture (top) and description structure (bottom). The architecture is divided into 2 main parts: the off-line encoding part (left arrow) and the run-time query part (right arrow). The MPEG-7 encoder extracts manually or semi-automatically the content features and returns an XML file according to the adopted description structure. The run-time query part starts from a request done by the end-user on a Java-based interface according to use cases specifications. The search capacities include management metadata (title, video summary, director, producer, production year), spoken words recognition, face recognition, noise and music presence, motion activity in the scene, female/male voice and semantic-based key frames search.

According to the data description structure, the whole movie is first segmented into visual shots. Each shot is further segmented in music, silence, speech or other audio content. The visual information in each shot is also segmented according to few representative key frames or in terms of face recognition and motion activity. The main encoding algorithms are described in the following.

4. ENCODING/RETRIEVAL MODULES

4.1 Visual features

Figure 2 shows a high-level summary of the main visual features encoding techniques used in MADIS. At this time, only the OCR module has not been implemented.

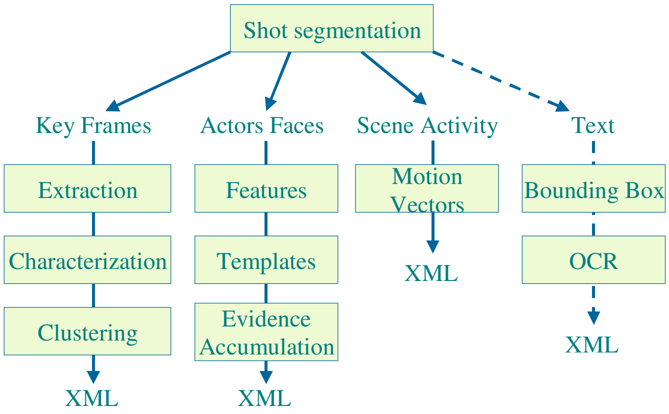


Figure 2: High-level block diagram showing the main visual features detection techniques used in MADIS

Key frames identification and clustering

For each shot, few representative key frames are detected based on the algorithm proposed par Vermaak et al. [14]. This approach allows to automatically find the number of optimal key frames and their position. The optimal key frames are defined as those that individually carry the most visual information and which are maximally different from the other frames in the shot.

The method is based on a representation of each frame in the shot by a low-dimensional feature vector. Chrominance representation is used in order to achieve some degree of invariance to illumination changes. The chrominance image is dividing in 2x2 regions from which the normalized red and green histograms are calculated. The full feature vector is composed of the frame number, the normalized histograms and the normalized average luminance of all pixels in the frame.

A Frame Utility Function" (FUF) is defined to measure the goodness of a frame as a key frame, in terms of the feature vector and the entropy of the color distribution in the image. This is based on the postulate that good frames are the ones that are well illuminated and carry the most information (measured from the entropy). The utility function assigns a low value to frames that are poorly illuminated. A Frame Distance Function (FDF) is also defined to capture the similarity between any two frames in the shot. This is again based on the above feature vector as well as the Bhattacharyya distance. Finally, a global utility function based on FUF and FDF is defined. The key frame sequence that maximizes this utility function is obtained by a non-iterative dynamic programming procedure.

The next step characterizes the key frames. This is done using low-level visual features from the XM and VRW libraries (color structure, dominant color, edge histogram, color layout, etc.).

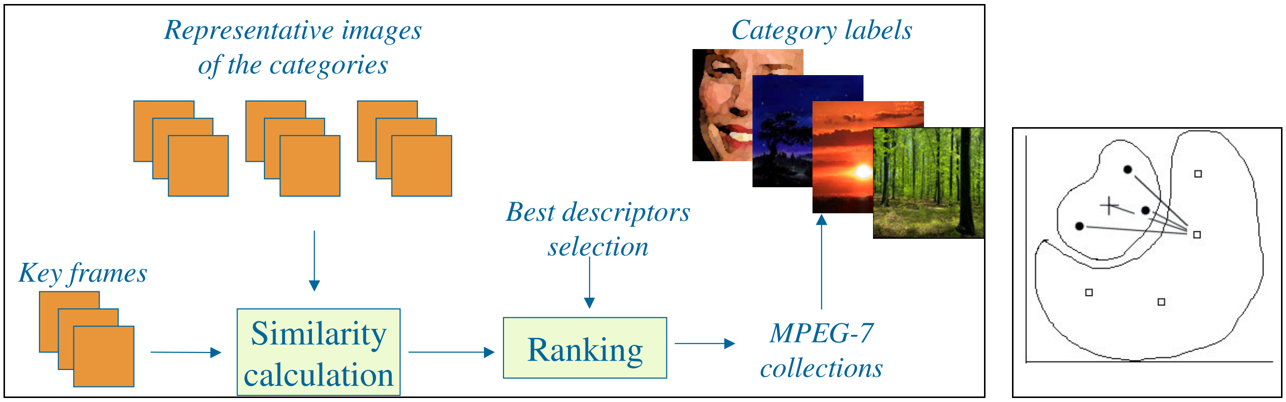


Figure 3: Block diagram showing the key frame clustering procedure (left).
Sketch illustrating the similarity calculation (right)

Finally, the whole key frame set is clustered into different categories (Figure 3). To each searchable category is assigned a "label" which is a typical image that represents the category. The approach is based on the hypothesis described on the right-hand side of Figure 3. We suppose that content-based key frame retrieval corresponds in identifying those (squares on Figure 3) that are close to a set of images having a similar semantic content (black points in Figure 3). The (unknown) class center (cross on Figure 3) of the category is replaced by a set of images (labels) having a human-invariant semantic content. The distance between a key frame and a category is the average between all the distances of this key frame to the entire representative images (black points in Figure 3) of the category.

The set of labels constitute a kind of "catalogue" that the user can use to search a particular set of key frames that correspond to a "human-invariant" semantic content: e.g. face, sunset, bridge, etc. The user can select only among those typical images in the catalogue. This means that the system does not allow a user to present an image, encode it and find the similar key frames. Such "open" search is long and often gives a lot of false results because of the human-dependent semantic interpretation of the input image. The main advantages of key frames search by category are that the "catalogue" is adapted to the type of database and the search is fast because there is no on-line similarity calculation. One can always increase the number of image in the catalogue once the database gets richer. A catalogue guides the external user to the categories available in the database and prevents searching for stuff that is no present in the database. The difficulty of this approach is however to identify the best descriptors that allow clustering of the key frames.

Face recognition

Face detection is based on the method proposed by Viola and Jones [16] and extended by Lienhart and Maydt [18]. It combines local Haar-like features decomposition with boosted cascade of simple classifiers (Gentle Adaboost). We have extended the approach [19] by the additional use spatial likelihood, of eye- and mouth-based template matching as well as Dempster-Shafer evidential theory [20]. On each face candidate detected by the cascade of classifiers, we attribute a spatial likelihood as function of the number of hits given by the classifiers (Figure 4). This allows us to define a face detection probability. The detection is reinforced using eye and mouth detection.

Face recognition is based on an embedded Hidden Markov Model (HMM) of the face proposed by Nefia et al., which uses observation vectors composed of 2D Discrete Cosine Transform (DCT) coefficients [17]. Using 2D DCT coefficients instead of pixel values reduces dramatically the size of the observation vectors and decreases calculation complexity. Each actor in the film is represented by an embedded HMM. A set of images representing different examples of the same face is used for training. After extracting the observation vectors corresponding to the test face images, the probability of the observation sequence is computed. Those probabilities are post-processed using the Dempster-Shafer evidential theory in order to identify a confidence level on the presence of the actor's face. This has been shown to provide better decisions and robustness against false alarm compared to the standard Bayesian approach [19].

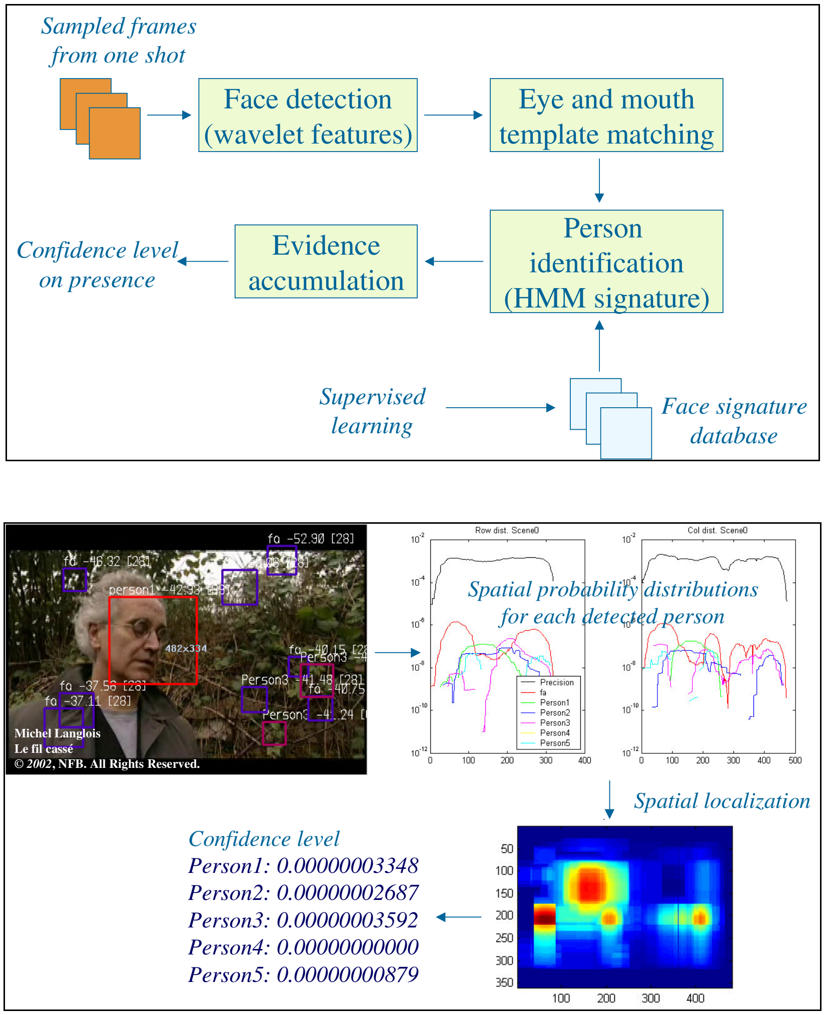


Figure 4: High-level block diagram showing the face recognition process (top).
Example of multiple detection processed by the Dempster-Shafer evidential theory (bottom)

The proposed approach has however some limitations/drawbacks. It is based on a structural type face detector (Haar-type basis), that is, color information is not used. There is no relation used between successive frames and it requires at least 100 detection to produce a reliable decision, which is a problem for very short shots. The processing is global on the whole scene, which is a problem for long, and complex shots with camera motions. Also, face recognition is limited to about +/-45 ° and the encoding time can be long. Among the possible ways to address those limitations, we plan to (1) take into account the face color information within a tracking algorithm (e.g. particle filtering) in order to improve face detection robustness to various poses, (2) use a transition probability between frames (Markov process) and (3) include an automatic face clustering (key faces) based on the color information or HMM encoding to make the process more automatic.

Motion activity

The motion activity descriptor has been adapted and corrected from the one included in the XM library. Details about the original algorithm are given in [21]. The approach is based on the displacement information of the macro-blocks in an MPEG-1/2 encoding. The speed variance is calculated for each P frames and averaged over the whole shot (Figure 5). An index ranging from 1 (low activity motion) to 5 (high activity motion) is then generated to quantify the variance.

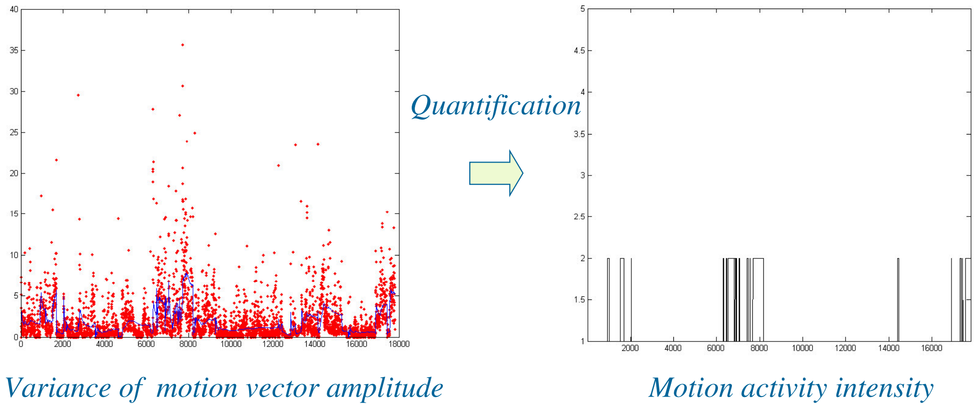


Figure 5: Example of motion variance and quantization result

4.2 Audio segmentation

Figure 6 shows a high-level block diagram of the audio encoding process. Audio classification is essentially a pattern recognition problem in which there are two basic issues, feature selection and classification. For feature selection, an effective representation should be able to capture the most significant properties of audio sound, robust under various circumstances and general enough to describe various sound classes. Among the numerous features proposed by researchers, e.g. [22], we choose to use the Zero-Crossing Rate, Spectral Flux, Low Energy and Linear Predictive Coefficients.

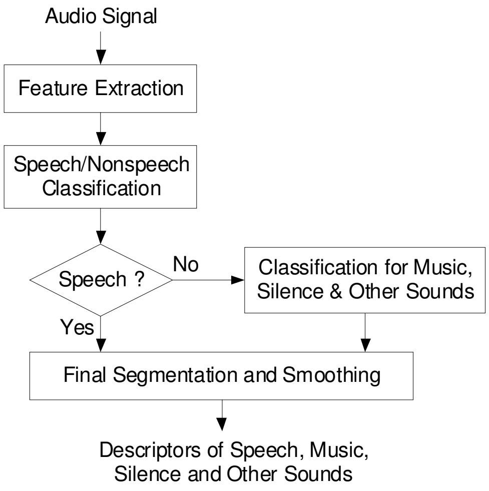

Audio Signal
Feature Extraction
Speech/Nonspeech Classification
Speech ?
No
Classification for Music, Silence & Other Sounds
Yes
Final Segmentation and Smoothing
Descriptors of Speech, Music, Silence and Other Sounds

Figure 6: High-level block diagram showing the main audio encoding process

For classification, the formulation of a distance measure and the rule of classification are critical. Before the audio sound can be classified, a codebook must be created. This is a supervised operation and requires the training data to be labeled; i.e. each training example must be associated with a class (e.g. speech or music). The elements of the codebook are the quantified data (i.e. the feature parameters) which form the feature space of the different classes. In fact, the feature space is partitioned into regions which have maximally different class populations. In order to reduce the error rate of classification caused by the overlap of the probability distribution of the features, multi-features (vectors) are used to classify the audio signal. Vector quantization techniques (e.g. Lloyd algorithm) are used to generate the codebook.

An audio file is discriminated by finding into which region of the codebook the input feature vectors are most likely to fall. Here, we use the N-Nearest-Neighbor algorithm to classify the audio signal by measuring the distance between the signal vector and the elements in the codebook. Four classes of audio signals have been identified in this project, speech, music, silence and other sounds. Finally, the classified information is reexamined by a smoothing algorithm based on the global decision, so that the burst error of classification can be remarkably reduced.

4.3 Speech recognition

While MPEG7 supports combined word and phoneme lattices, we choose to use phoneme-only recognition. It has been shown that phoneme-based retrieval can be as effective as word-based retrieval [23], but more importantly it solves the out-of-vocabulary word problem in recognition and unknown query terms in retrieval.

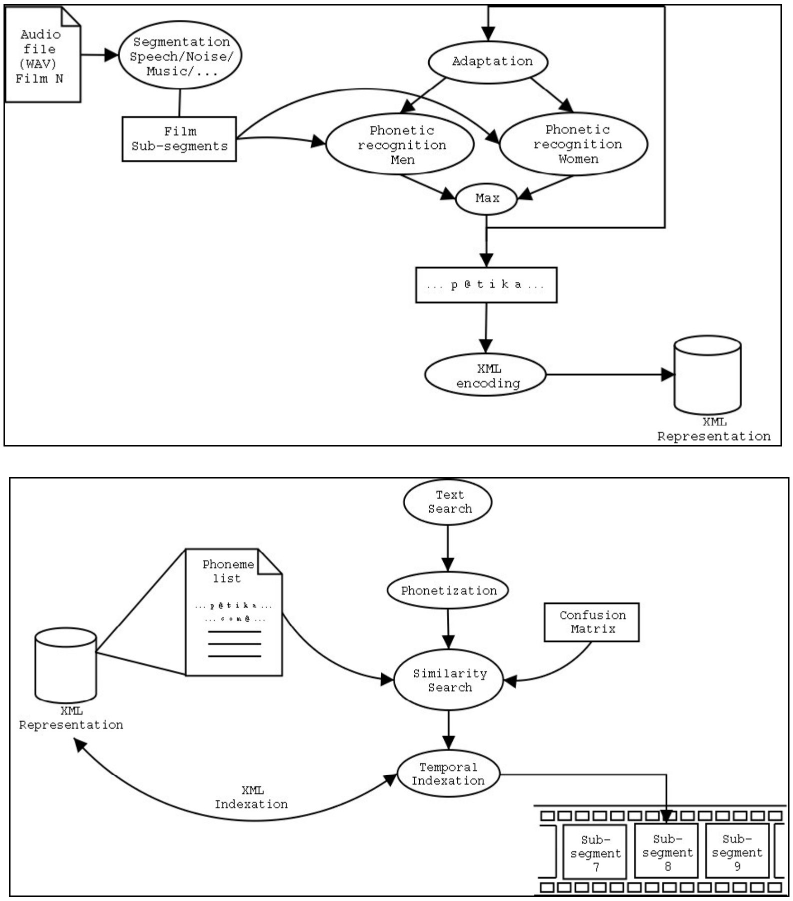


Figure 7: High-level block diagrams showing the encoding (top) and query (bottom) procedures for speech recognition

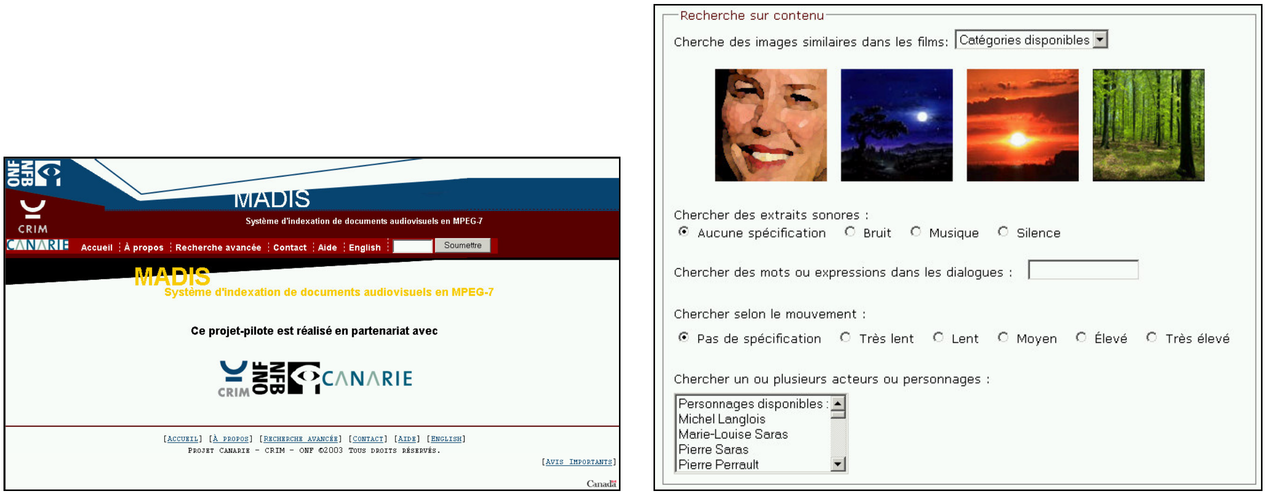
Our phoneme recognizer uses gender-dependent acoustic HMM, trained on a large database of Quebec French texts [13]. Each HMM models a phoneme given its left and right phonetic contexts. Phoneme sequences are modeled with an N-gram language model trained on representative texts, which have been mapped to phoneme sequences beforehand. We combine all the statistical models into a single recognition network, in the form of a weighted Finite State Transducer (FST). The recognizer processes each speech segment by extracting Mel-Frequency Cepstral Coefficients (MFCC) to form a sequence of input feature vectors. It then finds the path in the recognition network which maximizes the joint probability of the statistical models and the input feature sequence. To make the search efficient, we use a specialized version of transducer composition and a left-to-right beam search.

The first recognition pass is done with both male and female models in parallel, and the result with the highest likelihood identifies the speaker gender (top part of Figure 7). We take this best result and perform a gender-dependent, unsupervised adaptation of the acoustic models using Maximum Likelihood Linear Regression (MLLR). Using the updated models, we do a second recognition pass, combine all the second pass results for an entire film (separately for each gender), and perform another unsupervised adaptation using MLLR, followed by a Maximum a Posteriori (MAP) adaptation. Finally, the updated models are used to get a final, third recognition phoneme sequence for each segment.

The user enters its query in text form, which is first transformed into a phoneme sequence (bottom part of Figure 7). The similarity between this input sequence and recognized phoneme sequences is measured by a Levenstein edit distance that allows for insertions, substitutions and deletions. We derived edit weights from the recognition error statistics as recorded in a confusion matrix. FST operators combine query, edit distance and recognized sequences to produce a list of N-best matches, ordered by distance.

4. ON-LINE SEARCH DEMO

An on-line demo has been developed to experiment the query and streaming modules of MADIS from a Web-based application that will soon be available to the NFB Cineroute members on CA*net4 (Figure 8). CA*net4, as did its predecessor CA*net3, interconnects the provincial research networks, and through them universities, research centers, government research laboratories, schools, and other eligible sites, both with each other and with international peer networks. Through a series of point-to-point optical wavelengths, most of which are provisioned at OC-192 (10 Gbps) speeds, CA*net4 yields a total initial network capacity of between four and eight times that of CA*net3.


MADIS
Système d'indexation de documents audiovisuels en MPEG-7
CRIM
CANARIE
Accueil
À propos
Recherche avancée
Contact
Aide
English
Soumettre
MADIS
Système d'indexation de documents audiovisuels en MPEG-7
CRIM
NFB ONF
CANARIE
[ACCUEIL] [À PROPOS] [RECHERCHE AVANCÉE] [CONTACT] [AIDE] [ENGLISH]
[AVIS IMPORTANTS]
Canadä

Ce projet-pilote est réalisé en partenariat avec

PROJET CANARIE - CRIM - ONF ©2003 TOUS DROITS RÉSERVÉS.

Recherche sur contenu

Cherche des images similaires dans les films: Catégories disponibles

Chercher des extraits sonores :
● Aucune spécification ○ Bruit ○ Musique ○ Silence

Chercher des mots ou expressions dans les dialogues :

Chercher selon le mouvement :
● Pas de spécification ○ Très lent ○ Lent ○ Moyen ○ Élevé ○ Très élevé

Chercher un ou plusieurs acteurs ou personnages :
Personnages disponibles :
Michel Langlois
Marie-Louise Saras
Pierre Saras
Pierre Perrault

Figure 8: Snapshots of the current French version of the Web-based search demo of MADIS

Demo database is composed of 9 film excerpts produced by the NFB between 1971 and 2002. The film content is representative of typical documentary archives: indoor/outdoor and night/day scenes, interviews, music, speech, noise, faces, etc. Examples of requests that can be done include searching for (1) video segments done by "Producer" in "Year" where one sees "Person" saying "Word", (2) images of "Scene" in films talking about "Subject", (3) video segments of “Person” doing “Fast Motion”, (4) audio segments of film “Title” in which one hears music. The query result is a list of pointers to the beginning of key frame, visual, audio or speech segments that are found by the system to match the query.

5. CONCLUSION

We have reported about the development status of a modular and scalable MPEG-7 Audio-visual Documentation Indexing System (MADIS) with search capacity packaged into a demonstrable test-bed. The outcome of the project will be a Web site integrated within the Canadian National Film Board (NFB) Cineroute site and running over CA*net, a national Internet research and education network deployed by the Canadian Internet development organization CANARIE. This will allow end users to experiment search and streaming of audiovisual documents that have been indexed in MPEG-7.

Automatic content-based encoding of films is still an open problem in the R&D community. It is a formidable task to develop a system that is able to fully "understand" audiovisual content like a human does. The topic is so complex that researchers still largely concentrate on few specific aspects of it: still images, audio, natural scenes, sports events, manual encoding, without network capacity, etc. The MPEG consortium has compiled 43 international projects related to MPEG-7 in 2002 but only few of them address the whole audiovisual "picture" [10]. This is what MADIS aims at by addressing full MPEG-7 compliance, encoding and search, combination of three modalities (audio, speech and visual) and search capability on the Internet.

The development strategy for MADIS permits to address "hidden" scientific/technical issues that are very difficult to predict in such generic application. In order to minimize the technical risk, we have targeted a specific end-user application, that is, audiovisual documentary indexing. Use cases have been closely analyzed before the feature selection specification and the interface design. Audiovisual features extraction tools were selected and adapted from CRIM heritage or state-of-the-art open source codes.

Current limitations of MADIS include the followings: hard shot transitions detection only, no detailed semantics for image and video search (e.g. Quebec’s bridge, specific ship, specific animals, shot with a specific background, etc.), manual key frame clustering, no processing of speech with background noise, no music, noise or voice recognition and speech recognition in French only.

The next development phase of MADIS will include the following tasks: OCR encoding and search, automatic key frames and face detection clustering, query time reduction through the use of MPEG-7 collections, run-time performance tests, spoken word search accuracy, query results accuracy and final deployment (NFB Cineroute).

ACKNOWLEDGEMENTS

This work is supported in part by CANARIE Inc. (www.canarie.ca) under the Applied Research in Interactive Media (ARIM) funding program.

REFERENCES

1. C. G. M. Snoek, M. Worring, *Multimodal Video Indexing: A Review of the State-of-the-art*, Intelligent Sensory Information Systems, Univ. of Amsterdam, *ISIS technical report series*, Vol. 2001-20, December 2001
2. Y. Wang, Z. Liu, J. Huang, *Multimedia Content Analysis Using Both Audio and Visual Clues*, IEEE Signal Processing Magazine, Vol. 17, 12-36, 2000
3. R. Brunelle, O. Mich, C. M. Modena, *A Survey on the Automatic Indexing of Video Data*, Journal of Visual Communication and Image Representation, Vol. 10, 78-112, 1999
4. B. Furht, S. W. Smoliar, H. J. Zhang, *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers, Norwell, USA, 2th edition, 1996
5. A. A. Alatan, A. N. Akansu, W. Wolf, *Multimodal Dialogue Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing*, Multimedia Tools and Applications, Vol. 14, 137-151, 2001

6. N. Babaguchi, Y. Kawai, T. Kitahashi, *Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration*, IEEE Transactions on Multimedia, Vol. 4, 68-75, 2002
7. M. R. Naphade and T. S. Huang, *A Probabilistic Framework for Semantic Video Indexing, Filtering and Retrieval,* IEEE Transactions on Multimedia, Vol. 3, 141-151, 2001
8. See for instance RetrievalWare (www.convera.com); VideoLogger (www.virage.com); ImageMine (www.imageproducts.com) and Media Archive (www.tecmath.com)
9. *Introduction to MPEG-7: Multimedia Content Description Interface*, Edited by B. S. Manjunath, P. Salembier, T. Sikora, John Wiley & Sons, 2002
10. International Organization for Standard, *MPEG-7 Projects and Demos*, Draft document, March 2001
11. http://mpeg-industry.com
12. L. Gagnon, S. Foucher, V. Gouaillier, *ERIC7: An Experimental Tool for Content-Based Image Encoding and Retrieval under the MPEG-7 Standard*, Proceeding of the Winter International Symposium on Information and Communication Technologies (WISICT2004), Cancun, Mexico, Jan. 2004 (to appear)
13. G. Boulianne, J. Brousseau, P. Ouellet, P. Dumouchel, *French Large Vocabulary Recognition with Cross-Word Phonology Transducers*, Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), Istanbul, Turkey, June 5-9, 2000
14. J. Vermaak, P. Pérez, M. Gangnet, *Rapid Summarization and Browsing of Video Sequences*, BMVC'2000, 2000
15. P. Viola, M. Jones, *Robust real-time object detection*, Tech. Report No. CRL2001/01, Cambridge Research Laboratory, 2001
16. R. Lienhart, J. Maydt, *An Extended Set of Haar-like Features for Rapid Object Detection*, IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002; see also MRL Technical Report Dec. 2002 (http://www.lienhart.de/MRL-TR-May02-revised-Dec02.pdf)
17. V. Nefian, M. H. Hayes, *Face Recognition Using an Embedded HMM*, IEEE Conference on Audio and Video-based Biometric Person Authentication, pp. 19-24, March 1999
18. http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html
19. S. Foucher, L. Gagnon, *Semi-Automatic Actor Identification in Video Shots Using Dempster-Shafer Theory*, ICASSP04 (submitted)
20. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey, 1976
21. X. Sun, A. Divakaran, B. S. Manjunath, *A Motion Activity Descriptor and Its Extraction in the Compressed Domain*, IEEE Pacific-Rim Conference on Multimedia (PCM), LNCS 2195, pp. 450-453, 2001
22. E. Scheirer, M. Slaney, *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*, Proc. ICASSP 1997, pp. 1331-1334, 1997
23. K. Ng, *Subword-Based Approaches for Spoken Document Retrieval*, Ph.D. Thesis, MIT, Department of Electrical Engineering and Computer Science, February 2000, 187 pages