Toward an Application of Content-Based Video Indexing to Computer-Assisted Descriptive Video

L. Gagnon, S. Foucher, F. Laliberté, M. Lalonde, M. Beaulieu
R-D Department, CRIM, 550 Sherbrooke West, Suite 100, Montréal, QC, CANADA, H3A 1B9
{langis.gagnon, samuel.foucher, france.laliberte, marc.lalonde, mario.beaulieu}@crim.ca

Abstract

This paper presents the status of a project targeting the development of content-based video indexing tools, to assist a human in the generation of descriptive video for the hard of seeing people. We describe three main elements: (1) the video content that is pertinent for computer-assisted descriptive video, (2) the system dataflow, based on a light plug-in architecture of an open-source video processing software and (3) the first version of the plug-ins developed to date. Plugs-ins that are under development include shot transition detection, key-frames identification, key-face detection, key-text spotting, visual motion mapping, face recognition, facial characterization, story segmentation, gait/gesture characterization, key-place recognition, key-object spotting and image categorization. Some of these tools are adapted from our previous works on video surveillance, audio-visual speech recognition and content-based video indexing of documentary films. We do not focus on the algorithmic details in this paper neither on the global performance since the integration is done yet. We rather concentrate on discussing application issues of automatic descriptive video usability aspects.

1. Introduction

This paper presents the status of a project targeting the development of a prototype system to assist a human in the production of descriptive video. Descriptive video, also known as audiovision, is a narration added to the movie audio track which describes some visual elements to help blind and hard of seeing impaired people to fully enjoy their listening experience. According to the American Foundation for the Blind, there are approximately 10 million blinds and visually impaired people in the United States [1]. The Canadian National Institute for the Blind (CNIB) estimates this number to 1 million in Canada [2].

The descriptive video industry is planned to grow because of the imposition of federal obligations to broadcast more programs with descriptive narration. For instance, since December 2002, the Canadian Radio-Television and Telecommunication Commission (CRTC) made a requirement for new licenses issued to major conventional stations to broadcast a minimum of two hours a week of described Canadian television, rising to four hours by the end of the license period. This will increase, especially with the introduction of digital TV. Video description is still a relatively new service in Canada but is available in the United States since the early 1990's [1].

Computer-assisted tools for descriptive video require the development of algorithms that can automatically or semi-automatically extract visual content. This is closely related to the field of content-based video indexing. Development of content-based indexing systems is an active research topic. Among past or present initiatives, one can mention (1) the CIMWOS (Combined Image and Word Spotting) project [4], (2) the European network of excellence SCHEMA [5], (3) the VizIR project [6], (4) the Físhlár system, a suite of digital library applications which provides content-based video navigation services on broadcast video content over the Dublin City University campus [7] and (5) the open-source Caliph & Emir tool for photo annotation and retrieval [8]. There are also the IBM VideoAnnEx system [9], the Ricoh MovieTool [10] which assist authors to create video content descriptions interactively and the IBM MARVEL system [11] which has achieved the top performance on the TRECVID semantic concept detection evaluation in 2003 and 2004. Finally, one can mention a recent Canadian initiative, the MADIS project, which aimed at developing simple and practical MPEG-7 audio-visual content-based indexing tools for documentary films indexing and mining [12]. MADIS was addressing the global picture of content-based video indexing/retrieval with four requirements: (1) MPEG-7 compliance, (2) automatic encoding, (3) audio, speech and visual modalities and (4) Web-based search engine. The outcomes of MADIS provide the starting points of many indexing tools to be developed for the current project.

At this time, descriptive video is done off-line in the industry. Practical real-time video description for broadcast television does not exist yet, although a first tentative study has recently been explored by a Canadian group for Web-based applications [14]. However, adaptation of content-based indexing tools

is important because descriptive video use-cases are different. For content-based video retrieval, the user is typically a video archivist who is looking for a specific portion of the film containing a specific audio-visual content. In principle, there is no restriction on the visual content to be encoded (within the current technology's state-of-the-art) as they are all potentially interesting for video query. On the other side, video content to be described by a narrator can be quite limited by the content of the audio band (one cannot add an additional audio description when an actor is talking or when the sound level is high like in action scenes). On the other hand, there is no need to describe the visual information when it does not add useful insight to imagine what is going on in the scene. At present, there are no standard for descriptive video although some general guidelines exist ([15],[16]). For instance, one should describe what is essential to know such as [16]:

- actions and details that would confuse the audience if omitted
- actions and details that add to the understanding of personal appearance, setting, atmosphere, etc.
- any visible emotional states but not invisible information as mental state, reasoning, or motivation
- titles and credits and subtitles.

It is not our intention to discuss the validity of these guidelines or propose new ones. We rather look at the descriptive video problem through the "content-based indexing glasses" and try to figure out what would be a viable suite of indexing tools (within the video indexing state-of-the-art) that could help identifying visual content suitable for descriptive video, through a series of automatic screening processes. Of course, this is a purely scientific-driven standpoint that does not take into account all the industrial production constraints at this point. It has however the advantage of identifying feasible elements and a "wish list" of technological targets for future work.

The paper is organized as follow. Section 2 presents our view of the main visual content of interest for a computer-assisted multi-pass descriptive video system. Section 3 gives more details on the dataflow between these different concepts and the architecture of the tools under development. Section 4 briefly describes the current development status of some of these indexing tools. Note that we do not present the mathematical or algorithmic details of our tools in this paper. Most of them can be found elsewhere (see refs in Section 4). We rather focus on application issues related to automatic descriptive video and the status of our R-D activities within that context.

This work is part of the research theme "Audio-Visual Content Extraction and Interaction" of the new Canadian E-Inclusion Research Network [17]. The network partners are engaged in the development of audio-visual tools for the multimedia industry to improve the multimedia experience for the blind, deaf, hard of hearing and hard of seeing persons.

2. Visual content

For descriptive video (but also applicable to video mining), key-places, key-objects, key-text and key-persons are important elements to track and characterize at the different levels of a film structure.

It is generally accepted that a film can be decomposed in a hierarchical structure constituted of high- and low-level semantic parts. At the higher semantic level, a film is composed in "stories", or Logical Story Unit (LSU) [18]. A LSU is a collection of semantically related and temporally adjacent shots conveying a high-level concept in the same environment (place or locale) [19]. A shot is a consecutive sequence of frames recorded from a single camera. LSU boundaries are complex to identify as it is represented by discontinuities in space and time. Many algorithms have been proposed to tackle this task (e.g. [20],[21]). Shot boundaries are simpler to identify as they are a purely low-level visual process (hard transition, fade-in, fade-out, etc.). Very often, different LSUs share a common place. Some of those places are important for the story (key-places) [22]. Key-places are thus an important element for automatic place recognition, LSU segmentation and grouping.

Key-object conveys important meaning that is emphasized during film editing (e.g. a particular lamp that is important for the story). For practical indexing applications, object description must be robust to lighting conditions, scaling, blurring, partial occlusions, and viewpoint changes. An interesting approach is to describe the object with the help of affine invariant region detection [23]. In addition, some key-object can be associated to a key-place (e.g. a painting on a wall) and can also help to identify the place associated to a LSU.

Key-text is a text that is important for a human to understand the story flow (e.g. text close-up, transition text between shots, subtitles, newspaper title, street signs, ads). Locating any type of text or portions of text is still an active research area. Although many existing systems (academic or commercial) still concentrate on the extraction of caption text only, important progresses for unconstrained text detection in video has been achieved over the last ten years (e.g. [24],[25],[26]), often using a cascade of classifiers trained using Adaboost [27].

A person can be characterized by its body (e.g. clothes color, gait, gesture) and/or its face (e.g. mouth

shape, eye color, facial expression). In particular, for face and facial expression recognition, many automatic or semi-automatic recognition and interpretation algorithms have been developed over the last 15 years (e.g. [28],[29]). In almost all of these works, video acquisition conditions are strongly constrained by assuming, for instance, control over high-resolution images, good lighting conditions, faces frontal view of still person standing or sitting in front of the camera, etc. For commercial or documentary video applications, none of these simplifications can be assumed, resulting in a highly complex environment. We have recently proposed a new approach for face recognition in video based on the Dempster-Shafer theory to process accumulated statistical evidence of face presence within a shot [30]. For facial expressions in uncontrolled environment, an image-based approach based on Gabor wavelets is an interesting solution because they are relatively invariant to translation and illumination change and do not require the building of complex models [31].

Motion characterization describes the overall activity in a video sequence [32], either generated by the camera and/or objects motions. Techniques can be divided into two types: global and local. Global ones are based on motion estimation using optical flow or block matching combined to homographic transformations analysis between consecutives frames. Local ones involve foreground objects detection and tracking. Both types have to be considered for a complete understanding of video motion characterization (e.g. [33],[34]). An interesting architecture is described in [34] which is based on two modules. The first module is a low-level features extractor from which motion detection is based on the analysis of the homography transformation and the fundamental matrix between two consecutive frames. The detection of pan and tilt are computed through the homography transformation, while tracks and booms are detected through the epipolar geometry analysis of the images. The second module is a high-level feature extractor using segmentation algorithms to extract and label meaningful regions or objects in frames and tracking them.

Finally, another important concept is key-frame detection and categorization [35]. Key-frames are frames within a shot which visually summarize that shot. Key-frame categorization allows characterizing a shot in a global way (e.g. night or day, indoor or outdoor, dominant color, etc). The main issues in image categorization are the unknown number of categories, the high-dimensionality of the feature space, and the complexity of the natural cluster which are overlapping [36]. A popular way to find partitions in complex data is to use prototype-based clustering algorithms. The fuzzy version (Fuzzy C-Means) has been constantly improved for twenty years, by the use of the Mahalanobis distance, or the adjunction of a noise cluster. The competitive agglomeration proceeds by reducing the number of clusters over iterations to find the optimal number of clusters [37]. For supervised image categorization, images of a scene are represented by a collection of local regions (e.g. [38],[39]). Each region represents an instance of a bag corresponding to an image. The methods learn the instances distributions for each category.

3. Dataflow and architecture

The visual content depicted above and are dependent from each other and cannot be extracted from a single pass. For instance, key-object and actor face spotting requires the *a priori* knowledge on the object and visual examples for search and recognition. However, these are not strong technical constrains for off-line indexing. Each content calls for the development of a specific module with input coming from another module and output possibly feeding a third one. Figure 1 gives a block diagram of the dataflow between the various modules. We develop each module as a plug-in of the open-source video processing tool VirtualDub [40] in order to simplify the development process (especially for the common video I/O functionalities and the interface customization).

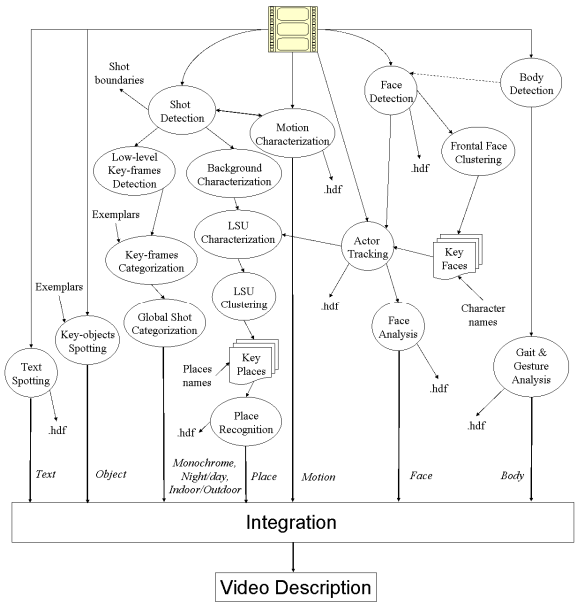

Figure 1: Dataflow diagram

Since video processing involves the management of a large amount of data, we have selected the Hierarchical Data Format (HDF) as the data exchange

framework between the modules [41]. HDF is a general purpose format for efficiently storing and retrieving heterogeneous scientific data such as images, arrays of vectors and grids. It has been developed by the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC). We use HDF5 for accessing intermediate processed data to help optimize the encoding algorithms and to efficiently transfer low-level data associated to visual content.

4. Indexing tools

A first version of the indexing modules has been implemented. Here, we present a brief description for five of them: (1) key-frame detection, (2) motion characterization, (3) key-face detection, (4) face recognition, (5) face characterization and (6) text spotting. All these tools run automatically at a processing rate of 2 to 30 fps on 640x480 videos (Pentium IV).

4.1 Key-frame detection

Key-frames detection is based on the similarity measures proposed in [35]. A key-frame is a frame which contains the most information and is the most distinct from the other frames within a shot. Two functions are used: the so-called "utility function", based on the entropy of the color distribution, and the "frame distance function", based on the Bhattacharyya distance. The information in the image is first compressed by dividing the chrominance image in 2x2 regions over which the histograms of the red and green components are calculated. Then, a feature vector for each frame is built from these histograms. Third, the utility function identifies the frames that carry the most visual information. Finally, the frame distance function quantifies the frame similarity. The user can select the number of key-frames he wants or let the system find the optimal number (Figure 2).

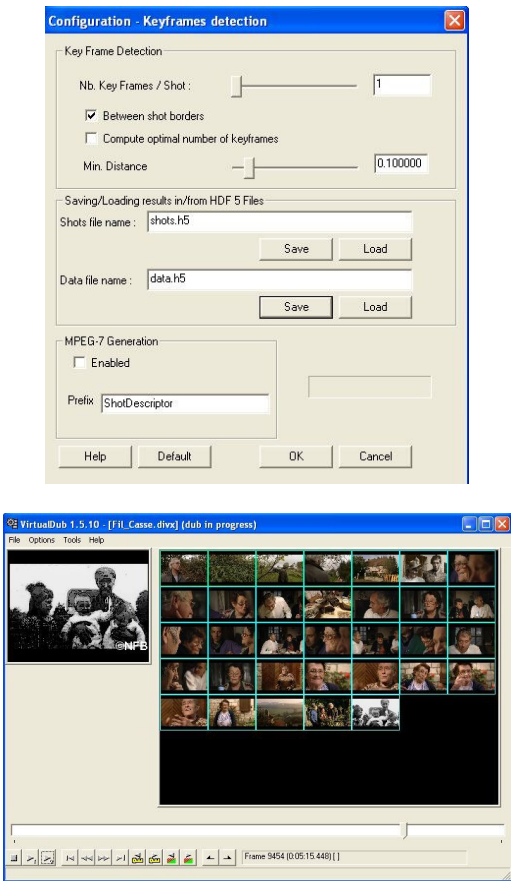

Figure 2: Key-frame interface for the selection of similarity distance threshold (top) and example of key-frames detection (bottom)

4.2 Motion characterization

Motion characterization is based on the Lucas-Kanade algorithm [42] which calculates the optic flow on automatically detected salient points (corners). At this time, motion characterization is global but a visual interface provides local information through intensity and direction motion maps (Figure 3). The grey map shows the motion intensity; color one is the direction. Image regions with a downward motion are labeled in green, left motion in blue, right motion in yellow and up motion in red. A polar histogram is also provided. In the first example on Figure 3, the person is moving to the left as the camera moves but with a lower intensity. The black regions are the ones where there is no motion. In the second example, the camera is zooming on the scene. The black region at the center is the zooming point. This plug-in is also used in another project within the E-inclusion research network to study the advantage of

positioning closed-captions in low motion activity regions for the deaf and hearing impaired people.

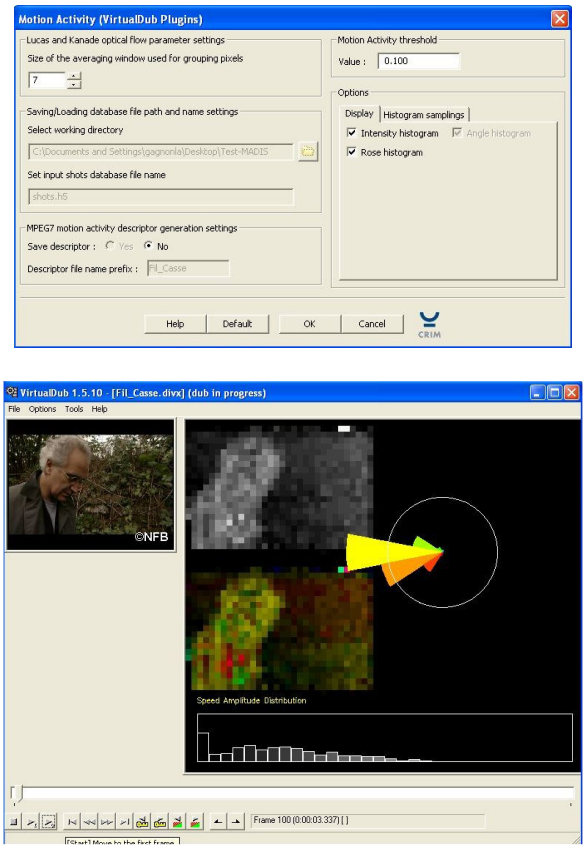

Figure 3: Motion activity plug-in interface (top) with an example of motion mapping (bottom). Left part shows the film frame. Right part show the motion intensity map (top), the motion direction map (middle), the polar motion direction histogram (right) and the motion intensity histogram (bottom)

4.3 Key-face detection

Key-faces are useful for automatic cast summarization [43] and automatic extraction of face image training sets for face recognition. Our key-face plug-in currently proceeds through two main steps: (1) frontal face detection and (2) face image dissimilarity assessment. First, face candidates are detected on each frame with a cascade of boosted classifiers [27]. A face frontal view selector, based on a spatial detection map of face candidates, is then applied. A frontal view quality factor $Q_{Face} = \#\{R_{Face}(x,y) = \max(R_{Face}) \mid (x,y) \in A\} / \#\{R_{Face}(x,y) = \max(R_{Face})\}$ is calculated which measures the fraction of maximal detections R_{Face} contained in a region A centered around the center of mass of the map maxima. A threshold is set on Q_{Face} to keep only frontal views (Figure 4).

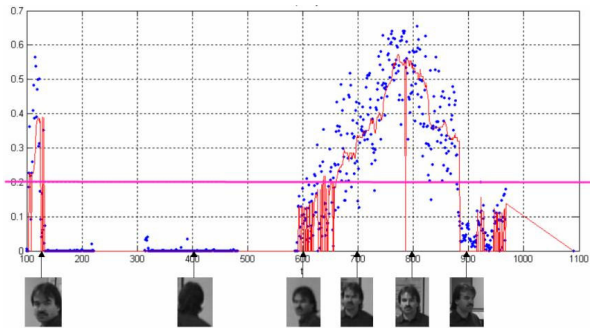

Figure 4: Example of frontal face quality factor values for various facial poses

Second, face image dissimilarity assessment checks if the currently processed frame contains a face image similar to the last detected image face. This step makes use of the same key-frame detection algorithm described in Section 4.1. As a result, only face images of the same person that are dissimilar enough are stored. This helps to provide a face image dataset with high face variability.

4.4 Face recognition

The face images dataset provided by the key-face detector are used as input for a face recognition module. (Currently, a manual refinement of the database is required.) Again, face candidates are detected on each frame with the use of a cascade of boosted classifiers. For each face candidate, a spatial likelihood is assigned to each face candidates detected. The likelihood is used as Dempster-Shafer evidential information to increase face recognition performance in complex scenes [30]. Face recognition of a face candidate is done with an HMM encoding procedure [44]. In our application, the HMM classifier takes a decision among a set of hypotheses which depend on the number of persons to recognize and a false alarm class is trained with non-face pictures. This identification process leads to a probability for each of the trained face (Figure 5).

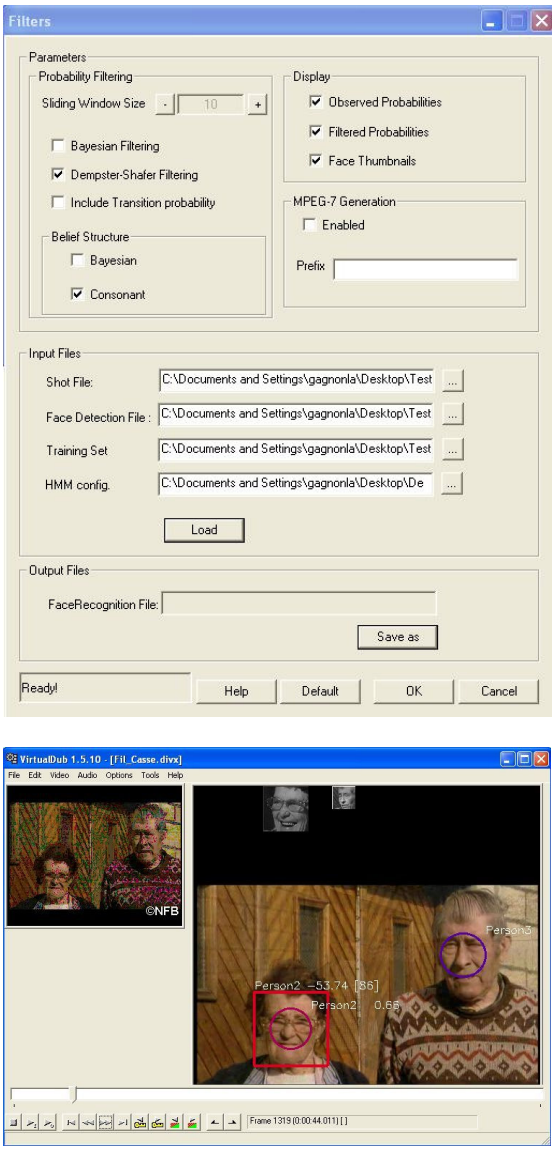

Figure 5: Face recognition plug-in interface (top) with a detection example (bottom). Red color means a high recognition confidence rate. Image thumbnails show the recognized faces. Thumbnail's size is proportional to the recognition confidence rate

4.5 Facial characterization

The face characterization plug-in provides information regarding the facial expression of almost frontal faces. The approach is similar to the key-face detector with three additional steps: (1) face feature detection (2) face normalization and (3) facial expression recognition.

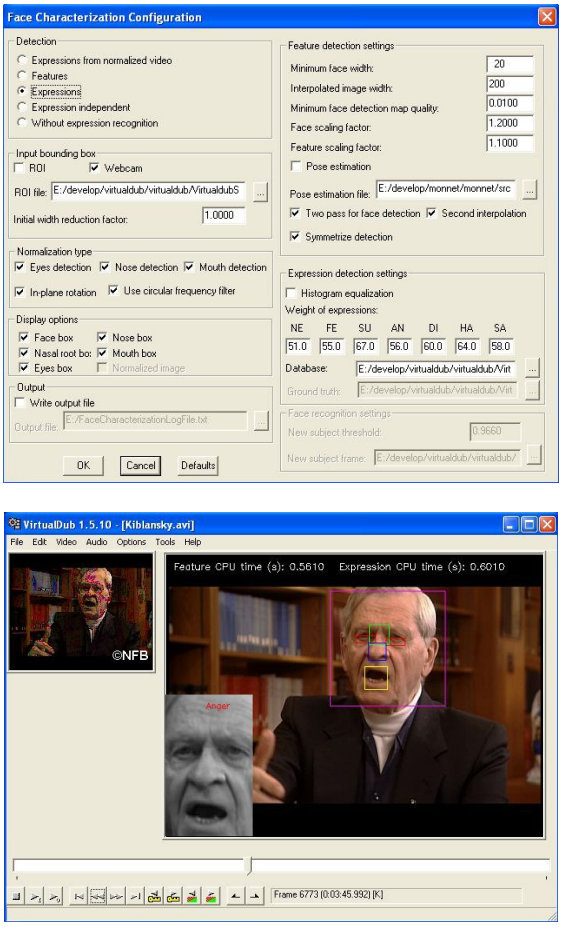

Figure 6: Face characterization interface (top) with detection example (bottom). The detected face points are given by the small squares. Lower-left corner shows the normalized image with the detected expression.

Facial features (eyes, nose root, nose and mouth) are detected using specialized boosted cascade of classifiers. Face normalization consists in using the position of these features to scale and orient the face image according to a normalized pose using affine, scaling, rotation or translation transforms. Gabor filters responses at different scales and orientations are then sampled on a regular grid covering the face to build a facial feature vector. The resulting vectors are normalized and reduced in dimensions before being fed to a classifier of seven basic emotions (anger, disgust, fear, happiness, sadness, neutral and surprise) (Figure 6) [45]. Emotions are detected globally on a frame by frame basis. Performance has been measured on a dataset of 2739 images coming from 7 public databases. The best facial expression interpretation rate obtained was 74.19% using a nearest neighbor classifier with a Euclidean distance similarity meas-

ure. The plug-in is a version of a face characterization module recently developed for a video monitoring system based on a set of loosely coupled cameras that build models and exchange visual information to track and recognize pedestrians [46].

4.6 Text spotting

The architecture is based on a cascade of classifiers trained using Adaboost [26]. Text spotting is performed by scanning the frame with a variable-size window (to account for scale) within which simple features (mean/variance of grayscale values and x/y derivatives) are extracted in various sub-areas. Training builds classifiers using the most discriminant spatial combinations of features for text detection.

Figure 7: Examples of text detection

In our implementation, we make no effort to constrain the learning algorithm to use predefined features, so the feature pool is very rich. The first stage is made of a simple classifier that discards as many false alarms as possible while letting potential candidates go through the following stages (with more complex classifiers). At this time, the output of the cascade is a decision map showing regions of interest that may contain text, and on which a commercial OCR will be applied (Figure 7). Performance was measured against a dataset of 147 key-frames extracted from 22 documentary films [47]. A detection rate of 97% is obtained with few false alarms.

5 Conclusion

We gave an overview of the status of a project regarding the development and integration of content-based video indexing tools to assist in the generation of audio description of visual information for the hard of seeing and blind peoples. We described three main elements: (1) the pertinent video content for descriptive video, (2) the chosen system architecture and (3) the indexing tools developed to date. Future works will concentrate on refining the tools presented above, improving their robustness and get final performance measures once they are all integrated (which is a lack of the present paper). In parallel, the development of other indexing tools like LSU segmentation, gait/gesture recognition and key-objects detection is currently under way. Integration in a global prototype solution for computer-assisted descriptive video is planned for the middle of 2007.

Acknowledgements

This work is supported by (1) the Department of Canadian Heritage (www.pch.gc.ca) through Canadian Culture Online, (2) the Natural Science and Engineering Research Council (NSERC) of Canada (www.nserc.ca) and (3) the Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE) of Gouvernement du Québec (www.mdeie.gouv.qc.ca). The authors thank J. Dutrisac from the National Film Board (NFB) of Canada (www.nfb.ca) for providing some of the video data.

References

[1] The Americal Federation of Blind: www.afb.org
[2] Canadian National Institute of Blind: www.cnib.ca
[3] Media Access Group at WGBH Boston: http://main.wgbh.org/wgbh/pages/mag
[4] CIMWOS project: www.xanthi.ilsp.gr/cimwos
[5] SCHEMA network of excellence: www.iti.gr/SCHEMA/index.html
[6] VIZIR project: http://vizir.ims.tuwien.ac.at/index.html
[7] Center for Digital Video Processing: www.cdvp.dcu.ie

[8] CALIPH & EMIR project: http://caliph-emir.sourceforge.net
[9] IBM VideoAnnEx project: www.research.ibm.com/VideoAnnEx
[10] Ricoh MovieTool project: www.ricoh.co.jp/src/multimedia/MovieTool
[11] IBM Marvel project: http://mp7.watson.ibm.com/marvel
[12] MADIS project: http://madis.crim.ca
[13] L. Gagnon, S. Foucher, V. Gouaillier, J. Brousseau, G. Boulianne, F. Osterrath, C. Chapdelaine, C. Brun, J. Dutrisac, F. St-Onge, B. Champagne, and X. Lu, "MPEG-7 Audio-Visual Indexing Test-Bed for Video Retrieval", IS&T/SPIE Electronic Imaging 2004: Internet Imaging V (SPIE #5304), pp. 319-329, 2003
[14] Canadian Network for Inclusive Cultural Exchange: http://cnice.utoronto.ca
[15] Guidelines for descriptive video: http://cnice.utoronto.ca/guidelines/video.php
[16] Guidelines for descriptive video: www.joeclark.org/access/description/ad-principles.html
[17] E-Inclusion Network: http://e-inclusion.crim.ca
[18] J. Vendrig, and M. Worring, "Systematic Evaluation of Logical Story Unit Segmentation", IEEE Trans. on Multimedia, Vol. 4, No. 4, pp. 492-499, 2002
[19] Handbook of Image & Video Processing, A.C. Bovik, (Ed.), New York: Academic Press, 2000
[20] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated High-level Movie Segmentation for Advanced Video Retrieval Systems," IEEE Trans. Circuits Syst. Video Technol., vol. 9, pp. 580–588, 1999
[21] Y. Rui, T. S. Huang, and S. Mehrotra, "Constructing table-of-content for videos," Multimedia Systems, Special Section on Video Libraries, Vol. 7, No. 5, pp. 359–368, 1999
[22] F Schaffalitzky, and A Zisserman, "Automated location matching in movies", Computer Vision and Image Understanding, Vol. 42, pp. 236, 264, 2003
[23] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Efficient Object Retrieval from Videos", 12th European Signal Processing Conference (EUSIPCO '04), 2004
[24] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, "VideoOCR: Indexing Digital News Libraries by Recognition of Superimposed Caption", ACM Journal of Multimedia Systems, Vol. 7, No. 5, pp. 385-395, 1999
[25] R. Lienhart, and A. Wernicke, "Localizing and Segmenting Text in Images and Videos", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No.4, pp. 256 -268, 2002
[26] X. Chen , and A. L. Yuille, "Detecting and Reading Text in Natural scenes", CVPR 2004, Vol. II, pp. 366-373, 2004
[27] P. Viola, and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", CVPR 2001, Vol. 1, pp. 511-518, 2001
[28] B. Fasel, and J. Luettin, "Automatic Facial Expression Analysis: A Survey", Pattern Recognition, vol. 36, no. 1, pp. 259-275, 2003
[29] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face Recognition Algorithms", Image and Vision Computing, vol. 16, no. 5, pp. 295-306, 1998
[30] S. Foucher, and L. Gagnon, "Face Recognition in Video using Dempster-Shafer Theory", ICASSP 2004
[31] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets", in Proc. International Conference on Automatic Face and Gesture Recognition, pp. 200-205, 1998
[32] S. Jeannin, and A. Divakaran, "MPEG-7 Visual Motion Descriptors," IEEE Tran. Circ. Sys. Video Tech., Vol. 11, pp. 720-724, 2001
[33] Y. Wang, T. Zhang, and D. Tretter, "Real Time Motion Analysis Toward Semantic Understanding of Video Content", CVCIP, Beijing, 2005
[34] Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. Ali, M. Shah, C. Grna, and R. Cucchiara, "Video Understanding and Content-Based Retrieval", Trecvid-2005
[35] J. Vermaak, P. Pérez, and M. Gangnet, "Rapid Summarization and Browsing of Video Sequences", BMVC 2000
[36] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and Semi-supervised Clustering: a Brief Survey", A review of Machine Learning Techniques for Processing Multimedia Content, 2004, 11 pages
[37] H. Frigui, and R. Krishnapuram, "Clustering by Competitive Agglomeration", Pattern Recognition, Vol. 30, No. 7, pp. 1109-1119, 1997
[38] P. Li Fei-Fei, and A. Perona, "Bayesian Hierarchical Model for Learning Natural Scene Categorie"., CVPR 2005
[39] Y. Chen, J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions", Journal of Machine Learning Research 5 (2004) 913-939
[40] VirtualDub project: www.virtualdub.org
[41] Hierarchical Data Format: http://hdf.ncsa.uiuc.edu/HDF5
[42] B. D. Lucas, and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", International Joint Conference on Artificial Intelligence, pp. 674-679, 1981
[43] A. W. Fitzgibbon, and A. Zisserman, "On Affine Invariant Clustering and Automatic Cast Listing in Movies", 7th European Conference on Computer Vision, 2002
[44] A. V. Nefian, and M. H. Hayes, "Face Recognition using an Embedded HMM", IEEE Conference on Audio and Video-based Biometric Person Authentication, pp. 19-24, 1999
[45] Facial annotation database deon by CRIM and publicly available for research purpose: www.crim.ca/fr/vis-projets.html
[46] L. Gagnon, F. Laliberté, S. Foucher, A. Branzan Albu, and D. Laurendeau "A System for Tracking and Recognizing Pedestrian Faces using a Network of Loosely Coupled Cameras", SPIE Defense & Security: Visual Information Processing XV, Orlando, April 2006, (to appear)
[47] M. Lalonde, and L. Gagnon, "Key-text Spotting in Documentary Videos using Adaboost", IS&T/SPIE Symposium on Electronic Imaging: Applications of Neural Networks and Machine Learning in Image Processing X (SPIE #6064B), San Jose, 2006, (to appear)