Univ Access Inf Soc (2009) 8:199–218
DOI 10.1007/s10209-008-0141-0

LONG PAPER

Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss

Langis Gagnon · Samuel Foucher · Maguelonne Heritier · Marc Lalonde · David Byrns · Claude Chapdelaine · James Turner · Suzanne Mathieu · Denis Laurendeau · Nath Tan Nguyen · Denis Ouellet

Published online: 5 February 2009

Abstract This paper presents the status of a R&D project targeting the development of computer-vision tools to assist humans in generating and rendering video description for people with vision loss. Three principal issues are discussed: (1) production practices, (2) needs of people with vision loss, and (3) current system design, core technologies and implementation. The paper provides the main conclusions of consultations with producers of video description regarding their practices and with end-users regarding their needs, as well as an analysis of described productions that lead to propose a video description typology. The current status of a prototype software is also presented (audio-vision manager) that uses many computer-vision technologies (shot transition detection, key-frame identification, key-face recognition, key-text spotting, visual motion, gait/gesture characterization, key-place identification, key-object spotting and image categorization) to automatically extract visual content, associate textual descriptions and add them to the audio track with a synthetic voice. A proof of concept is also briefly described for a first adaptive video description player which allows end users to select various levels of video description.

L. Gagnon (✉) · S. Foucher · M. Heritier · M. Lalonde · D. Byrns · C. Chapdelaine
R&D Department, Computer Research Institute of Montreal (CRIM), 550 Sherbrooke West, Suite 100, Montreal, QC H3A 1B9, Canada
e-mail: langis.gagnon@crim.ca

J. Turner · S. Mathieu
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montreal, QC H3C 3J7, Canada

D. Laurendeau · N. T. Nguyen · D. Ouellet
Department of Electrical and Computer Engineering, Laval University, Quebec, QC G1K 7P4, Canada

Keywords e-Accessibility · Video description · Video indexing · Computer vision

1 Introduction

This paper presents the status of a R&D project targeting the development of computer-vision software tools to assist the production of video description (also known as audio description, described video or audiovision). "Video description, or described video, consists of narrative descriptions of a program's key visual elements so that people with vision loss are able to form a mental picture of what is occurring on the screen" [1].

Just as closed captioning add visual information for the benefit of the hearing-impaired, video description adds verbal information on the audio track describing the visual content for the benefit of people with vision loss. The production process for both is quite long, but it is longer for video description; however, the time required can be reduced with the help of computer-assisted tools. At this time, video description is done manually by human describers and requires 20–30 h for a 1-h film, with a cost ranging from 2,500 to 6,000 $US per hour. The process typically requires a series of viewings, and validation of scripting and content, depending on the producer's protocol. The bottleneck is the video description scripting steps, which require numerous viewings in order to identify coarse and fine visual content elements and their importance for the story line. The first viewings involve a film summarization process to identify generic visual content such as shot transitions, whom the actors are and when they appear, where the action takes place, pertinent textual information, where the silent segments are in the audio track, and so on. This phase of the process is targeted in the

current project. The goal is to provide computer-vision tools for describers that can automatically detect generic visual content to assist during these first viewings. Such tools aim at reducing their workload and helping them produce more consistent content, more quickly, in order to increase the production quantity and ultimately the accessibility of media documents for people with vision loss. This application is timely, since the descriptive video industry is growing due to the imposition of regulations requiring broadcasters to add more descriptive narration in their programming. This will further be pushed by the deployment of digital television.

There is little literature, and no regulated standard of practice, regarding the production and usability aspects of video description [2–4], although some guidelines are available [5–8]. The content of the audio track can seriously limit the described visual content. For example, the describer cannot speak at the same time as an actor or when a silent segment is too short. On the other hand, there is no need to describe the visual content when it does not add useful information to the story line. Thus, the description should be limited to what is essential, such as [7]:

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

Not all people with vision loss appraise video description in the same way; user needs are diverse. Whatever the quality or quantity of the video description track, each individual has preferences depending on his or her level of vision, tastes, and experience. Thus, designing and implementing a player that is adaptive is an important consideration.

On the technical side, addressing computer-assisted video description requires the development of algorithms that can automatically or semi-automatically extract visual content, time-tag, and organize them according to the needs of the describer [9–12]. There has been little work in the field of computer-assisted video description. To the authors' knowledge, only the Live Describe project at Ryerson University has attempted to address this issue, and has provided interesting software tools to explore live video description [13]. In addition, it is clear that the development of computer-assisted video description tools is closely related to the field of content-based indexing and retrieval of digital video. Development of content-based indexing systems is a very active research topic, as demonstrated by increasing participation at the yearly TREC video retrieval evaluation (TRECVID) [14]. A literature review of the state of the art in video indexing could be the topic of a complete report by itself, and is well beyond the scope of this paper. A very recent state-of-the-art discussion of multimedia searching and indexing was published by the European Chorus Project Consortium [15]. Other recent related initiatives include (1) the combined image and word spotting (CIMWOS) project [16], (2) the European network of excellence in content-based semantic scene analysis and information retrieval (SCHEMA) [17], (3) the visual information retrieval (VizIR) project, supported by the Austrian Scientific Research Fund [18], (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 [19], (5) the Caliph and Emir tool for photo annotation and retrieval [20], (6) the IBM VideoAnnEx system for manual MPEG-7 indexing [21], (7) the Ricoh movie tool which assists authors in creating video content descriptions interactively [22], and (8) the IBM MARVEL system for semantic concept detection [23]. A recent Canadian initiative, the MADIS project [24–25], aimed at developing simple and practical MPEG-7 audio-visual content-based indexing tools for indexing and mining documentary films. MADIS addressed the global picture of content-based video indexing/retrieval and had four requirements: (1) MPEG-7 compliance, (2) automatic encoding, (3) audio, speech and visual applications, and (4) a Web-based search engine. The results of MADIS provided the starting point for many indexing tools being developed in the current project.

The work reported in this paper targets two user groups: describers and end users with vision loss. The needs of both are related to accessibility issues for this type of application: effectiveness of the describing process, and pertinence and fluency of the visual content description. Both are important elements of the adopted methodology, which is divided into three main parts:

- production practices: meetings with producers and analysis of described productions (films with a video description track) to get an understanding of practice within the industry,
- needs of people with vision loss: meetings with end-users to learn about their needs and to compare these with industry practice,
- system design and implementation: selecting and integrating computer-vision tools for automatic visual content extraction, taking into account the needs of describers and end users, as well as technical feasibility.

The rest of this paper is organized according to these three parts of the methodology. The first two sections provide details about the findings of the consultations with

the producers and end users, as well as from an analysis of a number of productions. These findings drive the project. The third section is more technical, and presents the current status of the proof-of-concept software tool, along with some performance results. The focus is not on the algorithmic details of the core technologies, but rather on the application issues of automated video description. The paper concludes with a summary of the achieved results, and of the work planned for the second phase of this project.

2 Production practices

2.1 Interviews with producers

This section presents the main findings regarding practices, obtained from consultations with three video description production companies. Interviews were conducted with seven people who manage or do video description. The consultations were taped, and subsequently transcribed. Each participant was asked to describe his or her work processes, and in particular what they considered most important, what was critical and how much time was required. Comments were then grouped and summarized by themes: production process, quality issues and general information on the video description industry. More details can be found in [26].

Typically, generating video description proceeds through the following steps:

- describers watch the film, and write a script describing key visual elements not discernible by listening to the dialog and sound effects,
- describers identify the silent parts in the audio track,
- describers carefully time the length of the description to fit within pauses in the dialogue and/or where there is silence,
- people with vision loss assess the quality of the results by reviewing the video description,
- a voice-over talent is cast, based on the genre of the program and the voices of the characters in the program,
- the video description is recorded,
- the description track and the original audio tracks are mixed,
- the finished audio product is transferred to the requested format for delivery.

The most time-consuming and complex process is the creation of the description itself. The producers reported in the interviews that "…at first, we do many viewings to assess (1) who is doing what and when, (2) what is the general idea conveyed by the film and (3) where the silences are, and how long they are". This work necessitates a lot of planning in order to get a general idea of how to create video description and where to place it. For a program series, general information on the production is kept so that various producers can access information that is already known, in order to foster consistent video description. For animated films, the process is often tedious. As one producer observed, "…we did animated films which involved many characters that had specific characteristics and that could transform themselves into other characters. This type of production requires a lot of planning to get a general idea on how to create video description and where to place it…there is always the risk of giving a description that is too literal, made of only pure information which could obfuscate what is interesting about the film". Another difficulty mentioned by producers is doing video description when there is a lot of dialogue. Description may well be needed, but there are few places in which to insert it. Some producers said that they are developing tools using voice synthesis to evaluate the number of words that can be entered into a silent space, depending on its duration; the describers then have to practice inserting the description with the right tempo, based on the space available on the audio track and on the type of film (which is different for a documentary than for an action film).

After the video description is generated, the production is usually viewed several times by a number of people before being approved. The objectives are to validate the consistency of the production, to verify the quality of the sound (according to the desired standard), to check the quality of the voice and the appropriateness of a male or female voice, and to ensure that names and other words are pronounced correctly, and that the vocabulary used is appropriate. Most producers have a list of criteria for accomplishing these tasks. In essence, "… we have high-level concerns such as: the names should be correct and pronounced properly, no incorrect information must be given about the action taking place, the speaking rate must be acceptable, and the quality of the sound has to be acceptable. If the content is abstract, then contextual information must be given. For example in animated art films where there is no dialogue but only music and visual effects, the technique chosen by the artist needs to be explained".

In addition, the producers were asked to list the kinds of visual content they most often describe, and this information helped identify the specifications for the system under development:

- face detection, including the frequency at which faces appear, in order to be able to identify characters correctly and consistently, so that this would serve as a good indexing tool,

- detecting segments of silence and their length, and the events surrounding the silence,
- building a player that allows end-users to repeat video description or get additional description,
- providing a summary of the film before planning production of the video description, in order to identify the characters in the film, and where and when they appear,
- describing the movements of people.

2.2 Analysis of described productions

An analysis was conducted of 11 productions of various types, in order to quantify the types of information found in the video description and the frequency of occurrence of each type, as well as to obtain other information. The productions included two feature-length films in French, of about 2 h each, two documentaries in English and French, and seven animated shorts in English and French, from the National Film Board of Canada (NFB) (Table 1). The main findings are summarized below; more details can be found in [27, 28].

The analysis is based on the typology of descriptions developed by Turner [3], refined by Turner and Colinet [4] and further refined in the course of this project (Table 2). Refinements included adding the categories "credits" and "video description". The "credits" category includes a former category covering information about the creation and dedication of a production, and the "video description" category covers information given about the video description itself, for example the name of the describing company, sometimes recited as part of the credits.

Overall, most of the information given in the video description in the productions studied in this project is in the following categories, given in descending order of occurrences: action (35–45% depending on the production), movement of the characters (7–22%), occupation, roles of the characters (3–18%), decor (4–12%), facial and corporal expressions (2–7%), textual information included in the image (1–2%), and Information about the attitude of characters (1–4%).

Other elements that were analyzed include the moment at which an episode of video description is recited in relation to the shot to which it refers, for example, at the

Table 2 Video description typology and global relative presence of each element in the productions analyzed

Video description typology	Presence (%)
Action	36–45
Information about the attitude of characters	1–4
Decor	4–12
Lighting	0–1
Spatial relationships between characters	1–3
Facial and corporal expressions	2–7
Clothing	1–3
Weather	0–1
Movement of the characters	7–22
Physical description of the characters	1–5
Indicators of proportions	0
Occupation, roles of the characters	3–18
Setting	7–9
Description of sound	0
Temporal indicators	1–3
Textual information included in the image	1–2
Appearance of titles	1–4
Credits	1–4
Video description	0–1

Table 1 Film productions with video description that have been analyzed

Film title	Producer	Year	Type
La vie est un long fleuve tranquille (film one in the text)	Étienne Chatiliez	1988	Feature-length
Le fabuleux destin d'Amélie Poulain (film two in the text)	Jean-Pierre Jeunet	2001	Feature-length
Voisins/Neighbours	Norman McLaren	1952	Animation
Il était une chaise/A Chairy Tale	Norman McLaren, Claude Jutra	1957	Animation
Le merle	Norman McLaren	1958	Animation
Blinkity Blank	Norman McLaren	1955	Animation
Caprices de Noël/Christmas Cracker	Grant Munro, Norman McLaren, Jeff Hale, Gerald Potterton	1963	Animation
Hen Hop	Norman McLaren	1942	Animation
En toute sécurité/Home Security	John Weldon	2004	Animation
Colonisation des plaines de l'Ouest/ Settlement of the Western Plains	Rex Tasker	1966	Documentary
Au pays de Riel/Riel Country	Martin Duckworth	1996	Documentary

same time as the shot, *N* shots before the shot to which it refers, or *N* shots after the shot to which it refers, as well as variations between the English and French versions. In the present study, in all films considered, over 85% of the time the episode is recited at the same time the corresponding shot is shown on the screen. Barely 5% of the time, the video description is recited while the previous shot is on screen, or begins then and finishes while the shot to which it refers is on screen. This is not typical of productions in general, according to the results obtained in previous studies. Of course, the recommended practice is to have the audience hear the video description at the time the corresponding shot is shown. However, this is not always possible because dialogues often occupy most of the sound space. In the productions studied in this project, the animated films have little dialogue, which explains the high percentage typical of this material.

3 Screenings and interviews with participants with vision loss

The aim of this phase of the work was to investigate how the various kinds of information found in the video description of the studied productions fit the needs of the user population. In collaboration with the Canadian National Institute for the Blind (CNIB) and the NFB, a number of screenings of these films with users who are blind or who have some vision loss were organized. Two screenings of feature-length films were organized by the CNIB. For each of these, about thirty participants were present. Four screenings were organized by NFB with smaller groups, in order to get their feedback. In addition, further screenings of films were organized from the NFB, with a 1-h program composed of two documentaries and seven animated films.

Following all these screenings, there was a discussion period, during which data were gathered on the reaction of the audience with vision loss to the quality of the video description. The issues addressed were how useful they found the description, whether it helped them follow the action, whether it encouraged them to see more films or not, what information they needed before viewing the film, what were the best and worst things about the video description, and what advice they could offer to producers towards improving the quality of the video description. The global results of the analysis of the data gathered correspond to that found in the scant literature available on the subject of video description. The technical and informational aspects provided more interesting results. In the following sections, a summary is presented of the findings gathered from screenings and consultations with more than fifty participants with vision loss. Details regarding the screening set-up, process, questions asked to the participants, the data collected from the audience and their specific comments are reported in [27] and [28]. In the following, only a qualitative overview of the findings is provided.

3.1 Technical aspects

People who are blind or who have some vision loss naturally focus their attention on the sound information when watching a film. For many participants, one of the cognitive difficulties is separating the video description recital from the sound track of the film. Sometimes this happens without problems, but other times, for example when there is narration, it is not always clear what parts of the information they are hearing are narration, and what parts are video description. One easy solution to this problem is that if the narrator is a male, the describer could be female. A substantial number of participants identified the volume of some sound, such as music at some points, as irritant, because it competed for their attention with the description information. Some sounds took up so much of the audio environment that it drowned out the video description, or made it hard for the participants to separate the two. On more than one occasion, participants said, “We’re blind, not deaf!”. When description is being heard, the volume of other sound needs to be lowered, so users can sort out the various sources of sound they are hearing. Essentially, the sound needs to be lowered from foreground to background. The complaint with the volume of other sounds was that it was so loud that it competed with the audio description being played at the same time, so that it interfered with the participants being able to decode easily the text of the description. Since the description is more important to them in understanding the film than the music is, the description should have priority in the sound space. However, participants also indicated that they liked the music, that it increased their enjoyment of the film and that at times it even was information-bearing in that it helped indicate mood or reinforced the action. They pointed out that they could still perceive this information if the sound level of the music were considerably below that of the description. Thus, music does not need to be eliminated or reduced to a large extent, but it should be lowered enough so that those who listen do not have to struggle to understand the text being recited. The phenomenon is akin to the cognitive task of trying to listen to the radio and converse with another person at the same time.

Describers should have a pleasant voice, as stated by the participants in this study, as this encourages the use of described materials. “Pleasant” means different things to different individuals: not disturbing, fluent and not robotic, with even intonation, etc. It also means “good articulation”

for some people, because when the video description is weak, the work of users in understanding what is going on is greatly increased. The participants also mentioned that "…the information given should not interpret the film, but rather describe it neutrally, all the while maintaining interest". This is somewhat different from what was recently reported in Fels et al. [29]. However, the findings of the present study do not claim universal consensus. The participants were not asked to compare different video description styles. However, a few different styles were presented to them, particularly in the art films of McLaren, in which the description is produced in such a way as to let the audience discover the art behind the animation. The description done by the NFB is given to make the film "understandable", yet it involves more than what a seeing viewer might get from watching the film.

There was a great deal of consensus on the part of the participants concerning the need to get some information about the film before seeing it. They said that some contextual information about the kind of film, a very brief summary, information about the genre, any technical particularities, and so on, should be given by the describer at the very beginning, before the film actually starts. Without this, they are suddenly bombarded with information and have no cognitive base on which to set it.

Some of the most important findings of the study are related to the need to tailor video description to individual differences. For any given film discussed in the study, some participants found that there was too much description, while others wanted more. In addition, some do not need to be informed about certain types of information, while others do. Many expressed the desire to have additional information at any point. While watching a film at home on DVD, for example, these people would like to stop the film at various points and hear more contextual information about what is going on, explanations of what has just happened, or additional information about the characters.

There is a broad range of both vision problems and individual differences, all of which should be addressed. Of those who avail themselves regularly of the services of the CNIB, fewer than 10% are totally blind. Most have some vision problem, and there are a great number of these problems. For each problem there are a number of levels of severity. Even users with the same type and degree of vision loss do not necessarily have the same preferences. In addition, there is a marked increase among the CNIB's clientele of users who are experiencing vision loss due to aging. With aging populations and increased longevity worldwide, the user base for video description can be expected to increase considerably in the coming years.

Because of this reality, the notion of flexible description is highly pertinent. Flexible description can be described as video description text marked up with information that could be used to help users set parameters or build a user profile. The main components would be identification, in each episode of description, of its level of importance in understanding the action and of the type of information it contains. Each episode of description could be coded with level 1, 2, or 3, for instance, and users could then decide which level they prefer. Each episode could also be coded with the type(s) of information it contains. The typology previously discussed could be used for this purpose. With the description thus coded, users who can see well enough to know that it is sunny in the picture, for example, would not need to hear such information. Users who can see the characters move could choose not to hear that information, and so on.

In summary, then, the important technical aspects have to do with potential conflicts between the description and the sound track, making the right choice of voice for the description, and the question of personalizing the level and type of description.

3.2 Informational aspects

The study participants pointed out that it is important to find the right balance between not enough and too much information. There needs to be enough information so that they can get an understanding of the action of the film, but not so much as to overwhelm them. They emphasize the cognitive difficulty of sorting and managing a lot of sound information when they are bombarded with it. They feel too much information is worse than not enough, because in the latter case they can sometimes manage to get some grasp of the action from the sound track or from what follows, whereas in the former case, they sometimes become exhausted and lose track of what is happening. The trick for describers is to get an understanding of what information can be deduced from the sound track, and what information needs to be stated explicitly.

Another problem stressed by the participants was the necessity of rapidly distinguishing a number of characters if several are introduced in a short period of time. When they do not succeed in doing this, they quickly lose track of what is happening in the film. This problem was particularly evident when screening the French film *La vie est un long fleuve tranquille* during the study. The film has many characters, and participants sometimes had difficulty sorting them out.

Describing where the action takes place is also important in providing the context in which the film unfolds. This kind of information needs to be provided so that users can understand in what kind of situation the characters are interacting. The corresponding visual information can often be understood instantly by sighted viewers, so the

challenge for describers is to find ways to provide this information quickly and economically. This is one motivation behind the development of the software tools presented in this paper. For instance, an automated tool provides time tags associated to the specific visual elements such as faces, places, and objects, which foster navigation of the film. This is an important factor in reducing the required time for conducting this process, as well as in searching and linking visual information in the film (e.g., key-places).

The important informational aspects, then, concern finding the right amount of description information and balancing it with that available from the sound track, allowing users to rapidly distinguish among the characters, and providing enough contextual information to indicate who, what, where, and when. Priority should be given to identifying the principal characters as soon as possible, then presenting the time, the place, and the action, all the while avoiding the pitfalls of interpreting the action and adding information not present in the image. This is the kind of contextual information that could be given before the film actually starts, akin to the tradition of scripts for plays, where there is a list of characters, a description of the time and place where the play is set, and information about where the first scene takes place, all before the first lines are recited or the action is described.

4 Software design and implementation

The production practices and end-user needs described in the previous sections are very informative as to the type and relative importance of content for video description, and the way it needs to be extracted. Automating the entire visual content extraction process, however, is far beyond the state of the art in computer-vision content indexing. At this point, a trade-off was made between what is technically feasible and what would be most useful to the producers, and aa system was developed that provides some film summarization elements that aim to reduce the producer's workload during the first steps of production of video description.

Up to now, work has concentrated on visual content, including key-place detection and clustering (important places that are recurrent in the film), key-face detection and clustering, key-text recognition and action (motion, gait and gesture). Detection of shot transitions and silent segments are also part of the developed system, but are not described here.

The current overall system architecture is shown in Fig. 1. The extraction of the audiovisual content is performed by a collection of specialized filters (the core technologies previously described) that extract high-level visual content. The overall processing is managed by the

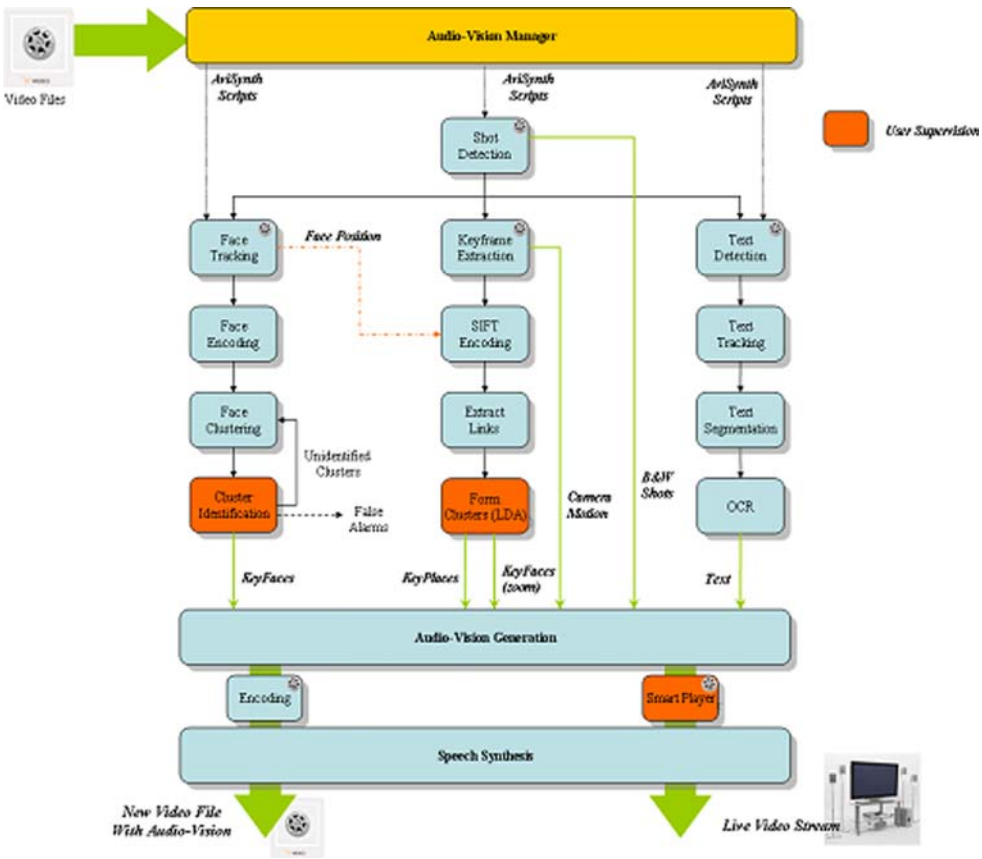

Fig. 1 High-level activity diagram of the full system. The film icon in the *top right corner of some boxes* indicates when a reading of the movie file is required

audio-vision manager (AVM), the primary goals of which are to ensure synchronization between the filters, acquisition of parameters, and generation of scripts for batch processing on a collection of video files. A final rough video description is produced by the audio-vision generator (AVG) with the goal of reducing redundancies by identifying logical video segments (LVSs), at the beginning of which the video description should be placed. Two end products are generated by the system:

1. A video file with a new audio band containing video description rendered by a speech synthesizer.
2. A video description script composed of sentences along with time stamps that can be passed to a speech synthesizer. This video description content can then be edited by the producer or directly read by an audio-vision player (AVP) along with the video.

The next sections describe the core visual technologies and their performance, as well as the AVM within which all the modules are implemented, and the audio-vision generator (AVG) which provides readable descriptions with timestamps attached. More details can be found in [26].

4.1 Core technologies

4.1.1 Key-places

At the higher semantic level, a film is composed of scenes called logical story units (LSU) [30]. A LSU is a collection of semantically related and temporally adjacent shots conveying a high-level concept in the same environment or place [31]. Some of these places are important for the story and are called key-places [32].

There is not much literature regarding classification of film shots in terms of key-places. Therefore, a new method was developed to find and cluster recurrent key-places in a movie [10]. The adopted approach is based on finding links between key-frames using probabilistic latent semantic analysis (pLSA) [33] to allow extracting and matching groups of local descriptors that may represent characteristic elements of a key-place. The pLSA generative models are used in natural language processing and statistical text analysis to discover topics in documents. They have been recently used in computer vision to solve various unsupervised classification problems [34–36]. An excellent presentation of the pLSA and LDA models applied to visual content can be found in [37].

In the pLSA framework, images are treated as documents, and the underlying goal is to extract topics (called visterms) that have semantic location characteristics (e.g., shelves of cigarettes from a bar, wallpaper from a room, etc.) using unsupervised learning. Visterm is a group of local descriptors matched with various images. The scale invariant feature transform (SIFT) [38] is used to extract local descriptors, followed by a Latent Dirichlet Allocation (LDA) approach [39] to identify the distribution of matches (topics) over the SIFT descriptors between key-frames. Visterms distribution is seen as part of a "topic",which is in fact a typical element representation from a scene at a higher semantic level. Figure 2 illustrates this approach. The whole process then follows four main steps: (1) key-frame extraction, (2) coarse linking using low-level image descriptors, (3) fine linking using latent aspect over SIFT matches, (4) final clustering using a spectral clustering approach [40]. Spectral clustering refers to a class of techniques which is dependent on the eigen-structure of an affinity matrix (i.e., the matrix of pair-wise affinities between points of the datasets) to partition the data objects into disjoint clusters.

The above method was tested on two full-length French movies of about 1.5 h each (*Le fabuleux destin d'Amélie Poulain* and *La vie est un long fleuve tranquille*). Figure 3 gives an example of links found for one location in the first

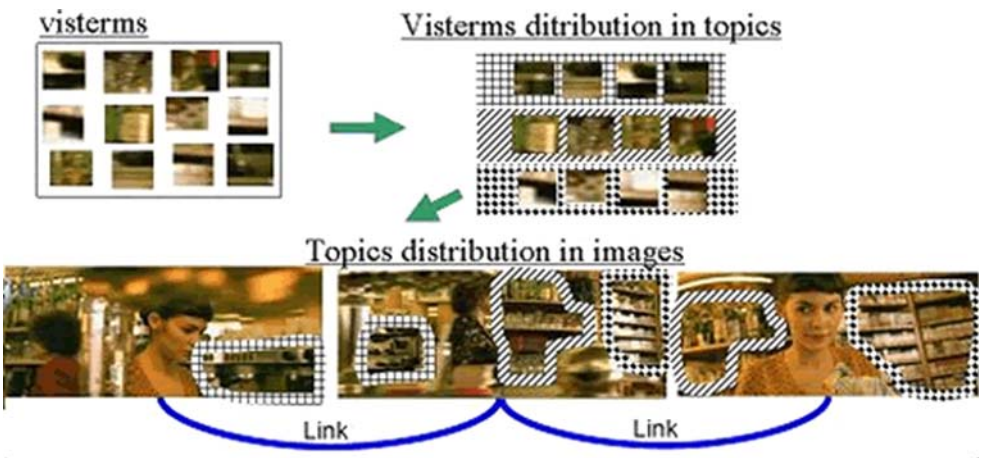

Fig. 2 Principle of latent semantic analysis

Fig. 3 Links for the grocery location. Matches are within white boxes

film. The visual content in each white box is the same (shelves) and is specific to the physical location. Even though this scene characteristic is seen from different points of view in different shots, the developed algorithm was able to recognize and link them, thus associating the four different images with the same place.

Tables 3 and 4 summarize the performance results regarding shot clustering as opposed to important locations in the film. Each film has four main locations appearing in 822 and 412 shots, respectively and that were manually identified. A location can be represented by several key-places (e.g., kitchen or bedroom for Amélie's apartment in film one). The algorithm automatically extracted a total of 131 shot clusters (key-places) in film one and 87 in film two. The recognition rate (RR) refers to the number of detected clusters representing a location. For example, the 32 detected clusters for the "bar" location represents 87% of the 239 bar shots. The top cluster RR is the rate at which the largest cluster has been assigned to the location. The false alarm rate (FAR) is the number of clusters assigned to the wrong location. The linking performance, i.e., the rate of correct linking among the shot set before the final spectral clustering, is 92% for film one and 96% for film two.

These results are quite encouraging, considering that locations in both films vary considerably because of the many points of view from which they are shot. In addition, the locations appear in several close-ups and in short shots. In these cases, there are fewer opportunities to extract their visual characteristics. The low FAR is also encouraging, because it is an important factor in minimizing the time the user needs to label the cluster.

Table 3 Evaluation measures for film one

Locations	# Shots	# Clusters (detected)	RR	Top cluster RR	FAR
Bar	239	32	0.87	0.47	0
Apartment 1	129	21	0.81	0.16	0
Apartment 2	94	22	0.68	0.27	0.01
Grocery	53	6	0.84	0.82	0
Total	822	131	0.78	0.47	0.01

RR is the number of detected clusters representing a location. *FAR* is the number of clusters assigned to the wrong location

Table 4 Evaluation measures for film two

Locations	# Shots	# Clusters (detected)	RR	Top cluster RR	FAR
Ground floor	117	24	0.70	0.15	0.02
Apartment	88	4	0.80	0.70	0.02
Garden	21	8	0.76	0.20	0
Office	16	5	0.44	0.13	0
Total	412	87	0.68	0.39	0.02

4.1.2 Key-faces

One of the main core technologies implemented in the AVM is the detection and time coding of the principal actors' faces (key-faces). A full-length movie typically contains a relatively small number of principal actors with many face instances and many faces of walk-on actors. Thus, the main challenge for key-face detection is to effectively cluster those faces among potentially thousands of face images.

Near-frontal view faces are detected and tracked in order to form trajectories (Fig. 4). Frontal-view face detection is performed using a cascade of weak classifiers [41], improved by a criterion that further rejects false alarms and faces that are not in a sufficiently frontal pose [42]. The faces detected are corrected for scale and translation to provide normalized face images where features are at the same positions in the image. The tracking algorithm is a particle filter (bootstrap filter) for which the likelihood of each particle is based on the response of a frontal view detector.

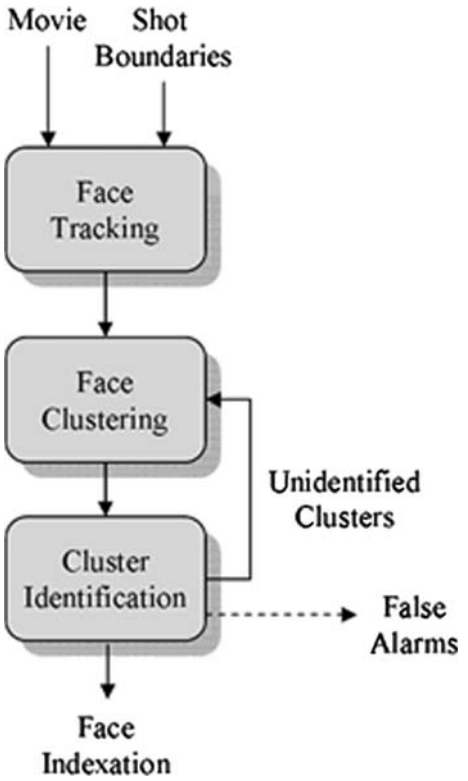

Fig. 4 High level description of the methodology for the indexation of faces

Face description is performed using a 2-dimensional principal component analysis (2DPCA) method [43–45]. The method has the following advantages over regular PCA (1-dimensional): (1) the size of the covariance matrix is either W × W or H × H for an image of H rows by W columns; (2) because the inputs are in fact the rows or the columns, the feature set is significantly enlarged, and this reduces the small sample size (SSS) problem; and (3) the 2D information is better preserved. It is noteworthy that the first two advantages mentioned here lead to faster and more robust eigenvector estimation.

Next, a spectral clustering algorithm [40] is applied in order to form clusters of similar faces based on the observed similarities between all the faces. The similarity measure is derived from the 2DPCA features computed on each face. The choice of the optimal number of clusters (a validation problem) is not critical. Over-clustering of the dataset is not an issue, as long as it produces several homogeneous clusters per face.

This approach was tested on the feature-length movie *Le fabuleux destin d'Amélie Poulain*, for which 14 actors had been previously identified manually. The tracking of detected faces resulted in 1,287 trajectories. For each trajectory, the three best faces were kept for training, and from this the 2DPCA eigenvectors were computed. The affinity matrix was then computed, and then the spectral clustering was applied with 1,000 iterations and 60 clusters. The best partition was retained, using the criterion of maximum average silhouette value. Figure 5 shows some of the clusters obtained.

The clustering result is presented to the user in the following way. Clusters are ranked according to their average cluster silhouette. Within each cluster, tracks are also ranked according to their decreasing silhouette value, so that outliers are displayed in last positions. A perfect cluster is a homogeneous cluster with no outliers that the user can label (i.e., associate the name of the person with the cluster) directly without further processing (Fig. 5). However, a cluster is still considered acceptable if the top tracks (largest silhouette values) are homogeneous. Figure 6 shows the results on cluster homogeneity for the above dataset. The cluster "purity" is the percentage of face thumbnails representing the same face in a cluster. Each point on the graph indicates the percentage of clusters having purity above a given threshold. For example, 100% of the clusters have purity at least greater than 20%, and 60% of the clusters have purity at least greater than 50%. Although no cluster is "pure", the algorithm pre-sorts a large amount of information, thus reducing the manual clustering and labeling workload of the user. When purity is low, the user can filter out faces manually, or repeat the process further on the unidentified tracks in an attempt to improve purity.

4.1.3 Key-text

Key-text is text that is important in understanding the story flow (e.g., text close-up, intertitles, subtitles, newspaper titles, street signs and credits). Locating and deciphering any type of text or portions of text is still an active research area. Although many existing systems (academic or commercial) concentrate on the extraction of caption text only, important progress in unconstrained text detection in video has been achieved over the last 10 years (e.g., [46–48].

The developed text-reading module is divided into three main components: (1) text detection, (2) text segmentation and (3) OCR. The text detector locates regions of interest (ROI) in the video frame that possibly include text. These ROI are then segmented before being processed by the recognition component linked to an OCR system. All this frame-based extracted information is temporally handled by a tracking component.

The text detection stage consists of scanning each video frame with a variable-size window (to account for scale) within which simple features (e.g., mean/variance of gray-scale values and x/y derivatives) are extracted using a cascade of classifiers. The best features have been found with a learning algorithm trained using Adaboost [47] that analyzed thousands of examples of pieces of images containing

Fig. 5 Examples of automatically generated face clusters

text or not. The result for each frame is a set of regions of interest where text is expected to be found. Typically, if the image contains a string of characters, the detection generates a chain of overlapping regions of interest. They are then aggregated to yield blocks of text (Fig. 7). Detection performance was measured against a dataset of 147 images extracted from 22 documentary films of the National Film Board (NFB) of Canada. A detection rate of 97% was obtained with relatively few false alarms [12]. The false alarms are filtered out later by a language model (see below).

The ROIs containing text are then segmented to remove background and noise. This is a difficult task since no information is available about the text color in relation to the background color. This difficulty is minimized by considering only the pixels that are in the region of the centroid of the ROI. The RGB values of these pixels are then collected, and the K-means clustering algorithm is invoked to find the three dominant colors (foreground, background and noise). Forcing the K-means algorithm to find three clusters instead of two improves the quality of the segmented characters (more detail is available in [12]).

Character recognition is performed by commercial OCR software. Each call to the OCR returns a set of character strings. More than one string may be associated with the ROI. A simple language model based on bi-grams is used to distinguish the meaningful strings from the useless ones. It uses frequencies of occurrence of pairs of letters collected from large French-language text databases to compute a score for each string. The "real" string is the one having the best score. The bi-gram frequency table was

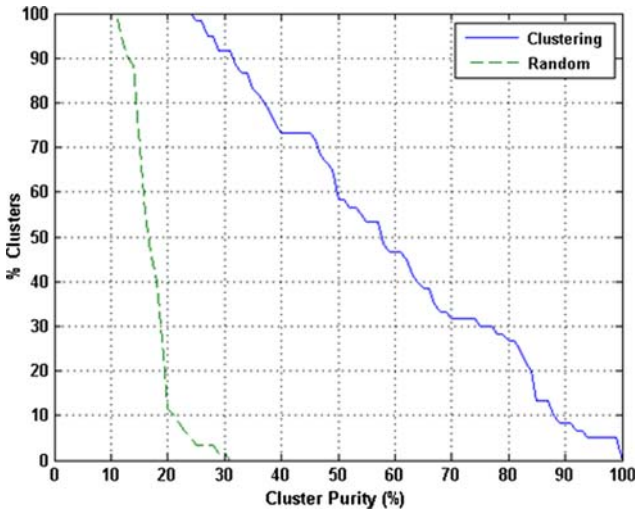

Fig. 6 Cluster purity (homogeneity) obtained for the film one. *Dash curve* is the result of a random cluster label assignment on thumbnails

computed from a set of 31 million characters drawn from various sources, including the French newspaper *Le Monde* [49].

Finally, tracking is necessary to aggregate the information in time, so that two pieces of the same text detected in successive frames are labeled only once. The tracking strategy is to aggregate text boxes based on their relative positions. This aggregation is called a track, and it holds information about the text being tracked, such as a unique identifier, the location and size of the corresponding ROI in the image, the frame number where the text first appeared or disappeared, the recognition results for each frame, etc. Typically, detection may yield a number of text boxes. When a given box is in the same location as a track, it is merged with that track (i.e., the track is extended and its data are updated), otherwise a new track is created. Tracks that have not been updated for a certain number of frames are terminated, and their information is stored.

As a track terminates, all the strings collected throughout the duration of the track are analyzed. The goal is to end up with a "representative" string for this track. String variability may be more or less pronounced, depending on the quality of the segmentation and the recognition. This is performed using a language model developed by the Speech Recognition team at CRIM that ranks strings according to their linguistic "quality". The model is based on statistics on word bi-grams (as opposed to the character bi-grams mentioned above) extracted from various sources.

4.1.4 *Human actions*

Recognizing high-level human actions (walking, running, sitting, gait, gesture, etc.) requires the detection and tracking of the motion of body parts. Two approaches are being explored to achieve this goal: detection and tracking of (1) low-level moving features and grouping into coherent clusters belonging to moving humans and (2) skin-colored regions (e.g., faces, hands) which are naturally associated with the presence of humans. In both cases, it is necessary to take into account the overall motion in the scene, in order to compensate for camera movement. Here, only the first approach is described (more details can be found in [50]).

The feature-tracking module currently being explored (Fig. 8) extracts motion information from the video sequence. It is based on the Kanade-Lucas-Tomasi (KLT) feature tracking algorithm [51–53], which selects features from an initial image and tracks them in the following images. The module provides local motion information for scattered points on a frame-to-frame basis. Once refined by clustering algorithms, the tracking information is used to track camera movement [54], locate the background scene, and finally, find the people or objects moving in the scene. The algorithm performs three main tasks:

- SelectGoodFeatures: useful features are located by examining the minimum eigenvalue of 2×2 gradient matrices.
- TrackFeatures: features are tracked using the Newton-Raphson method of minimizing the difference between

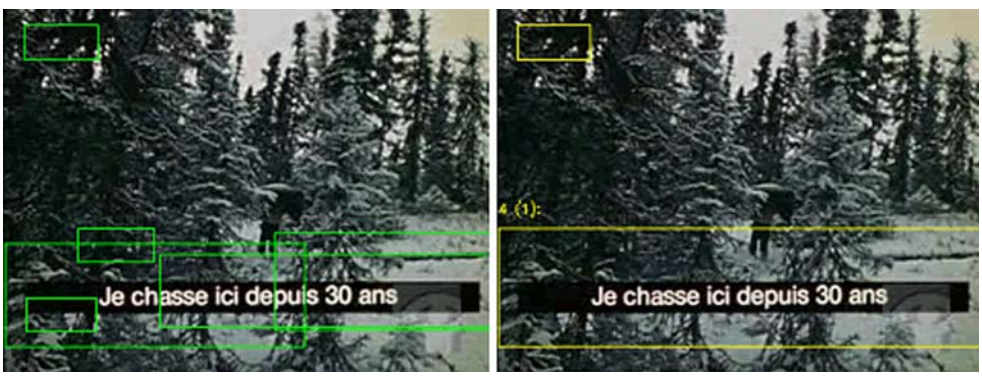

Fig. 7 *Left* basic text boxes hypothetically containing text. *Right* resulting boxes following aggregation

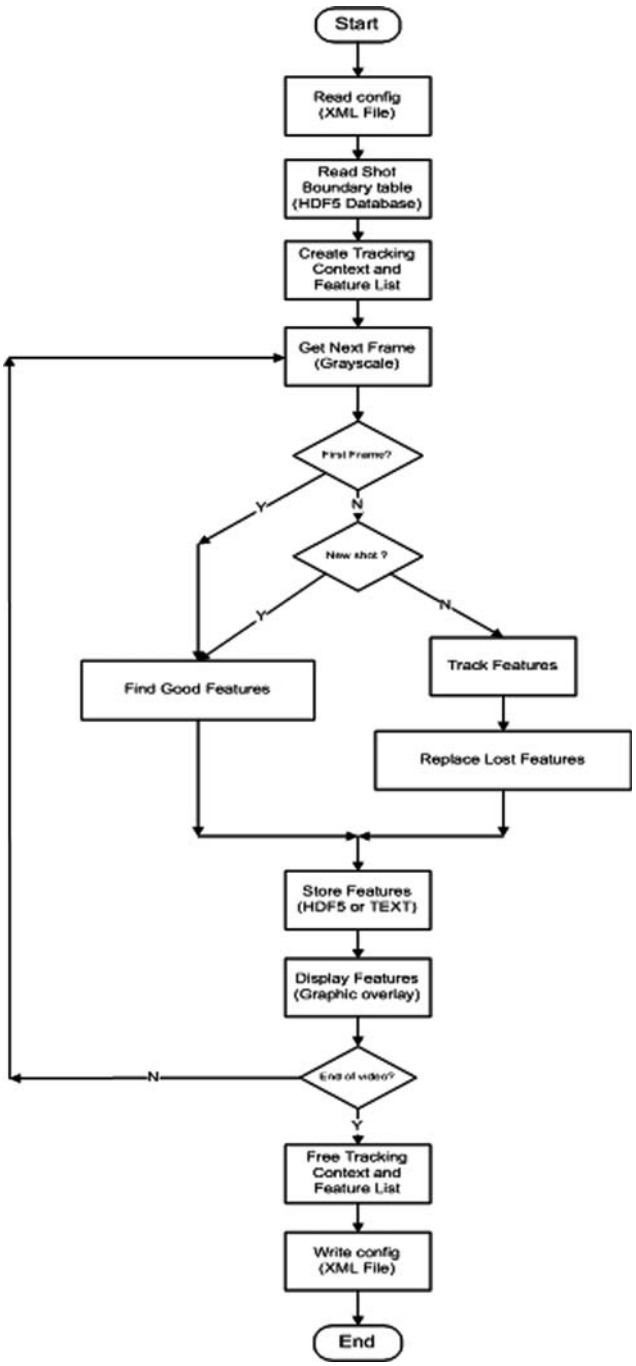

Fig. 8 KLT feature tracker flow chart

two windows. A multi-resolution approach helps the tracker to find important displacements between images.

- ReplaceLostFeatures: similar to FindGoodFeatures, but only lost features are replaced. For instance, the continuous replacement of lost features allows the discovery of new or occluded objects in the scene.

In addition, other functions are made available to support the various data structures such as the tracking context, the feature list and the feature history. In Fig. 9, some examples of tracking are presented for the film *Le fabuleux destin d'Amélie Poulain*. The KLT feature tracker is a powerful algorithm and relatively fast (2–10 frames/s depending on the frame dimension and number of feature points), considering that it is an optical flow approach. The motion information obtained by the KLT feature tracker is essential for the analysis of sequences with camera movements.

Background subtraction and motion detection are the basic tasks of motion analysis. The objective is to separate the foreground objects from the background scene in a video sequence with various camera movements. This involves the following sub-tasks:

- estimation of the camera movement,
- clustering based on frame-to-frame velocity of the KLT features [55] and separation of the background from the objects in the foreground,
- clustering of the objects in the foreground.

At this point, only shots in which the camera movement is limited to panning are considered. It is also assumed that the background motion dominates the movement of the objects within the frame. To estimate the camera movement [54], the system (1) calculates the average speed v_t of all features in each frame t and (2) applies a polynomial fitting to find the appropriate average speed of the camera from v_t throughout the shot. Figure 10 shows an example of result of the estimation of the camera movement. In this shot, the camera initially pans quickly to the right, then slows down, and stops at the end of the sequence. The velocity vector in the x direction is first negative, then becomes null near the end of the sequence, while the velocity vector in the y direction remains approximately null.

Motion clustering is performed frame by frame using a fuzzy C-mean (FCM) algorithm [56]. The vertical and horizontal velocity components of the features act as a discriminator for this clustering. Since clustering is performed independently on each frame, the label of the output categories can change, so a link of the labels across the shot needs to be performed. This can be achieved by matching the respective characteristics such as centroid distance, Hausdorff distance [57] and membership counts. This operation is necessary when the camera movement is not known. When an estimation of the camera movement is available, some of the categories with the background can be directly associated. The clusters identified as potential background then have their average speed matched with the camera movement.

Clustering of the foreground objects is achieved using a hierarchical clustering method. The object discriminator is the minimum Euclidian distance between two groups (closest points). Hierarchical clustering starts by grouping the closest features together, and then iteratively merges the groups. Figure 11 shows an example of a result. In the

Fig. 9 Examples of feature tracking results. *Straight lines* represent the feature history over the last five frames

example, the main foreground cluster is considered to be the one with the most members among the clusters.

4.2 Audio-vision manager

The AVM is a direct response to the needs of the producers as expressed in the Sect. 2, that is, the needs to have a software tool assist the generation of video description.

Each of the aforementioned visual content extraction technologies has been implemented as an AVIsynth [58] module that reads the video and produces raw video description information. Each one requires adjusting the parameters to achieve their goal, and some are dependent on the output of others. The AVM manages the processing and ensures proper synchronization among the modules; that is, the acquisition of parameters, the order of execution, the inter-dependencies among the modules, the integrity of the output data, and the generation of scripts for batch processing on a collection of video files.

The AVM software tool requires four directory paths as shown in the upper part of Fig. 12: (1) the binaries path where all modules are placed, (2) the videos path, (3) the configuration path which contains the parameters for each module and their execution scripts, and (4) the data files path where the output data produced by the modules are stored.

The AVM scans each directory, retrieves information on the binaries, parses the content of the data files, creates a video list and checks the configuration files. These operations are executed automatically at startup or when a path changes. For each module, a configuration file is created using the default values made available by the binary file. Feedback on the status of the system is provided to the user. For each module-video pair, a script file is created to execute all modules on all videos. A status field indicates that all scripts are ready to be executed. Finally, the progressive status of the data to be produced by each script is displayed (output datatype status field in Fig. 12). In this example, the column indicates that data files do not yet exist, and thus scripts should be executed. To be processed, each script in the left pane can be added to the batch list in the right pane. Then a single click on a button starts the batch processing of all selected scripts. Ultimately, all the modules are executed on each video file, providing the video description for it.

4.3 Audio-vision generator

The AVG module is the last step before the output can be sent to a user. At this point, the AVM has processed the video one or many times, and extracted raw visual information. The purpose of AVG is to convert this raw

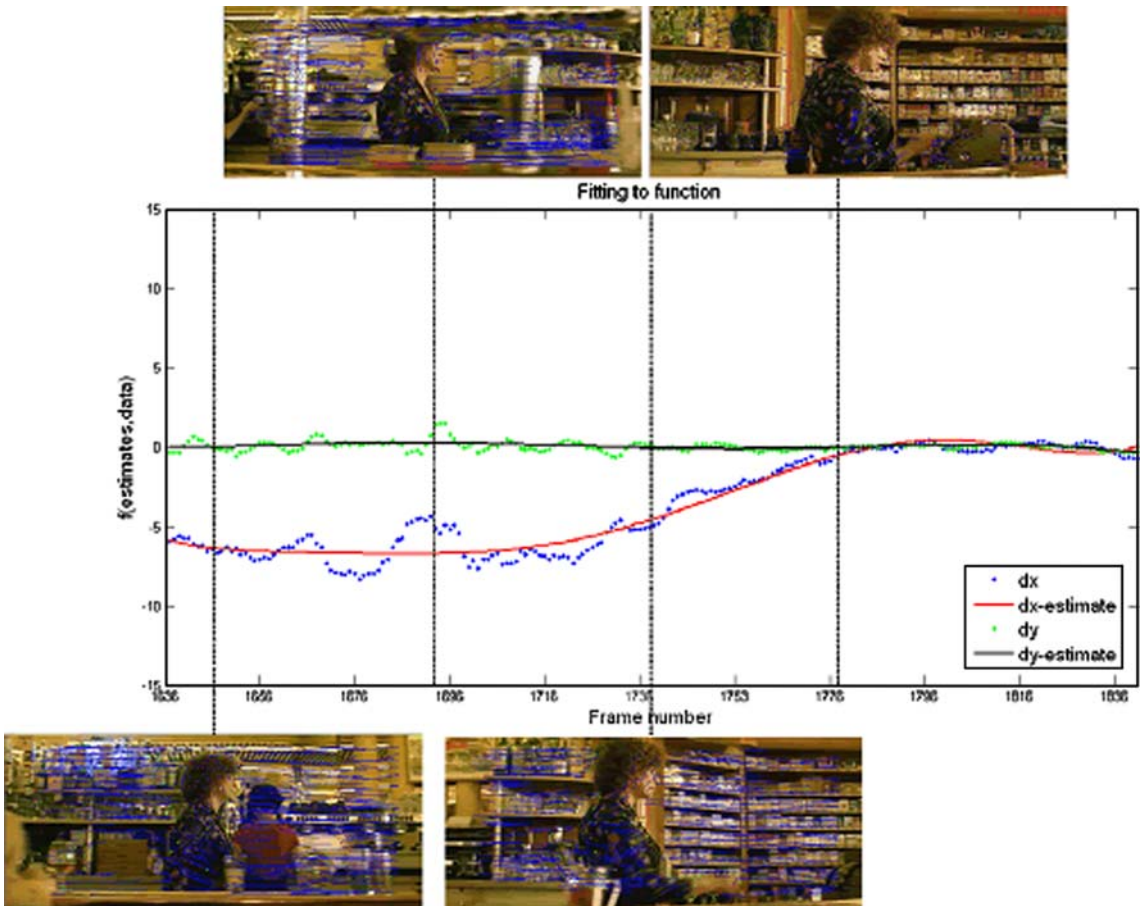

Fig. 10 Camera motion estimation. The *X*-axis is the frame number, the *Y*-axis represents the camera motion. The *top* and *bottom dotted curves* are the velocity value in the *y*- and *x*-direction, respectively

Fig. 11 Example of foreground object clustering. Original (*left*) and main foreground object (*right*)

information to readable descriptions, and to add time-stamps to them. In the end, this provides a comprehensive data file containing all the extracted video description.

To achieve this goal the AVG processes four distinct elements: (1) the raw information decoding, (2) the construction of three LVSs, (3) filling of these LVSs, and (4) the video description synthesis (Fig. 13).

First, the AVG needs a description input file. This is entered manually by the user, and lists all the text strings that could be added as video description in order to decode raw information. For example, each face detected by the face detection module has an associated cluster number representing an individual actor. The description file maps the cluster number to the name of the character.

The second step creates the LVSs of three different types: places, shots and audio segments. These LVS are used to store information from other modules that can be synthesized at the beginning of each LVS. For example,

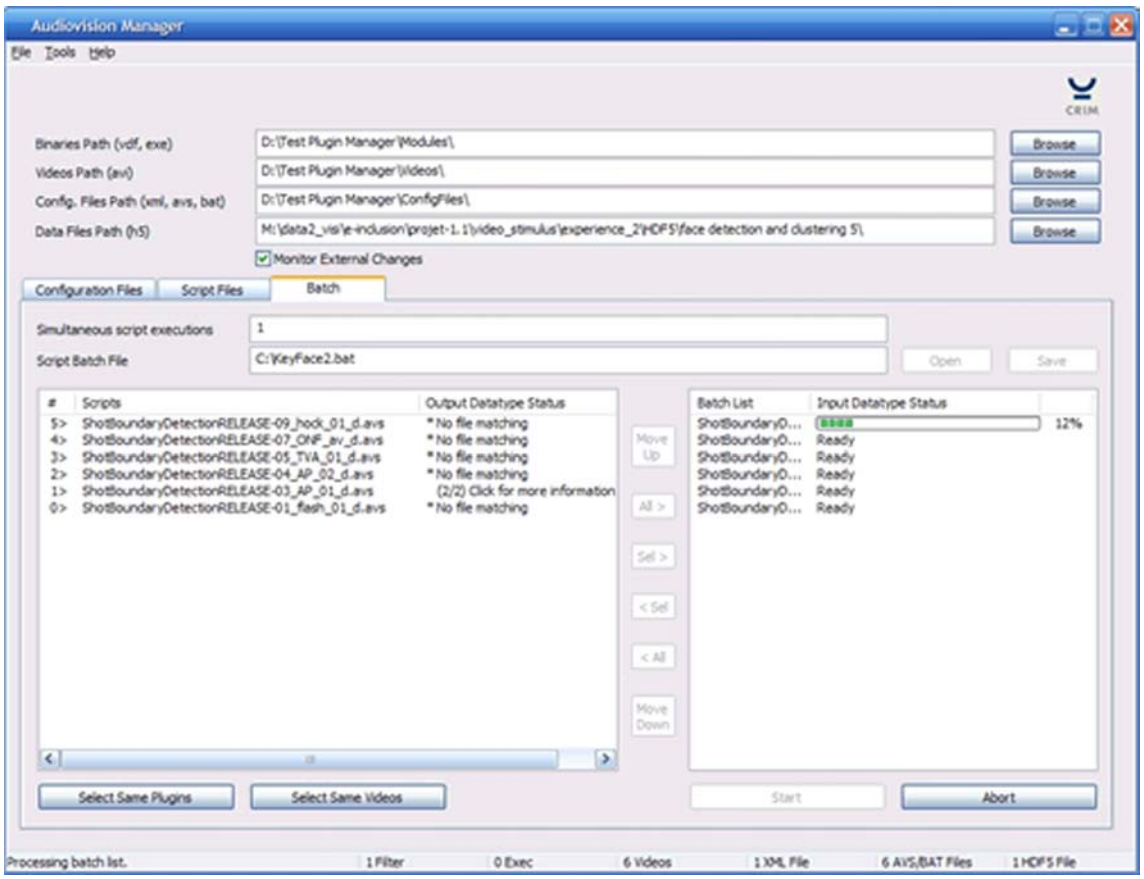

Fig. 12 Main window of the audio-vision manager. The AVM is the control and coordinate the calls to the visual content algorithms and provide time tags

five place-type LVSs can be seen in Fig. 13. Each description that can be associated to a place is stored in this type of unit. Another type of LVS is the shot boundary. Shorter than a place unit, the shot boundary is used mainly when no place is detected or when more than one description should have been added to a single place. Finally, the audio LVS is used to segment a position where a video description could be added without disturbing the understanding.

Third, all other information is added to an existing LVS. The motion detection is added to the closest shot boundary LVS determined by an automatic shot boundary detector (not described in this paper; see [26] for details); the motion type is therefore read at the beginning of the shot. Face detection is added, if available, to the place detection LVS, so that each character is named when a new place is detected. The text detection is added directly to the closest available frame from the audio segmentation, in order to avoid a delay between the moment when the text appears and the moment when it is read.

Finally, the last step merges the various LVSs into a single timeline, if necessary, to order amalgamate multiple descriptions from a single video frame into a sentence and to put a time stamp on each of them. Thus, all the video description that could be read during video playback is made available in a unique location, with an associated time stamp, ensuring that no dialogue is interrupted.

4.4 Audio-vision player

As a follow-up to the most important finding of the section entitled “Screenings and interviews with participants with vision loss”, namely the need to have a software tool that allows selecting a level of video description, a first prototype of an adaptive audio-vision player has been developed. The current version appears beside the main application during a video playback, and adds the possibility of changing the level of video description at any time. Since the video itself does not contain any video description in this case, the dynamically created video description recital is sent to the describer using an external speech synthesis program. Figure 14 shows the various controls available to the user during playback. In the upper part, the reading options determine the way video description is read. The reader can be changed and the

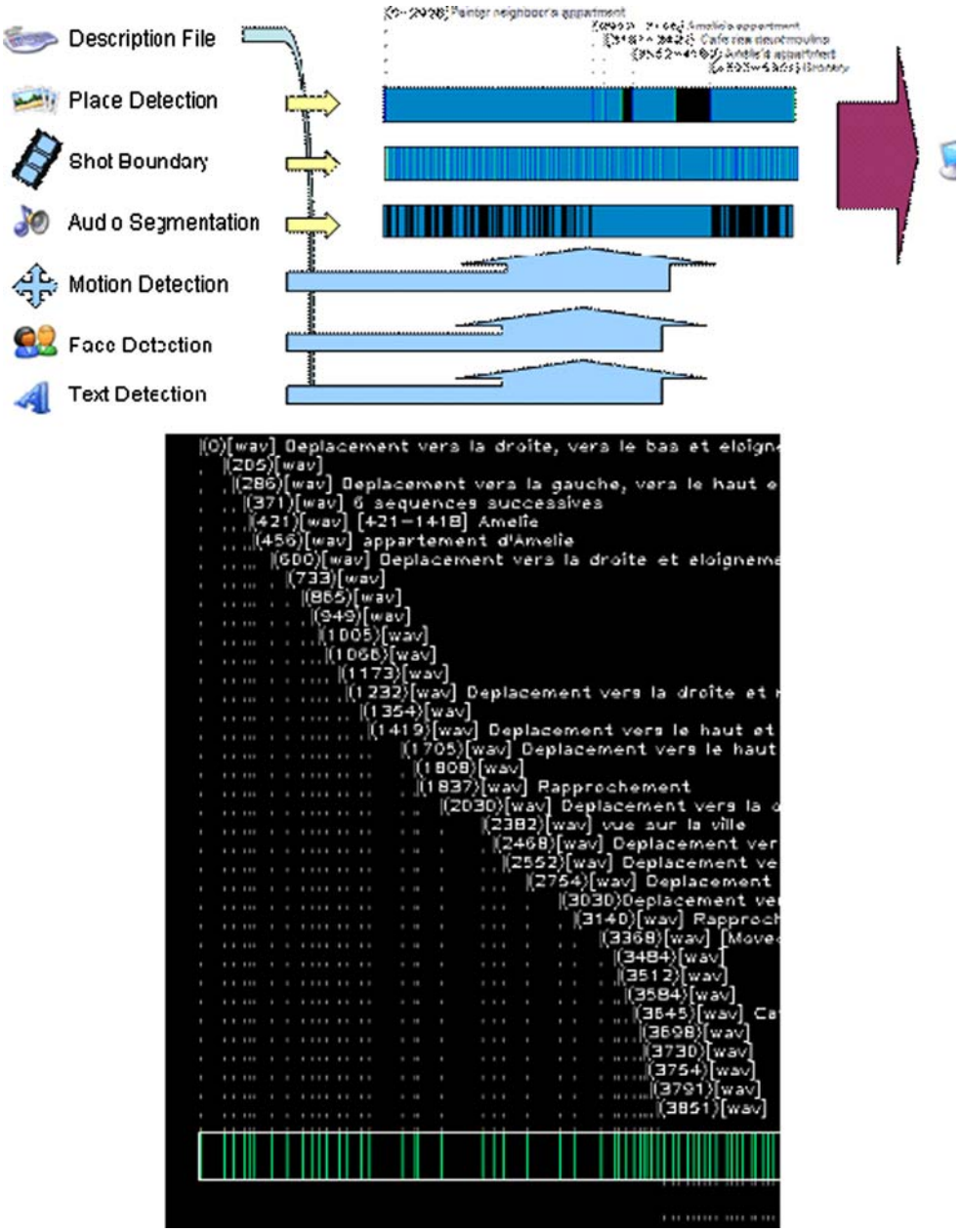

Fig. 13 Data merging process for the audio-vision generator (*top*) and example of readable descriptions with time tags associated to them (*bottom*). The later constitutes the output of the system that can be sent to the voice synthesizer and added on the audio band of the film

reading speed can be increased or decreased. The lower part of Fig. 14 shows the available video description layers that can be activated or not. The last two layers, the character identification and the action place, offer the additional possibility of being recalled at any time by clicking on the corresponding button.

A static version of the AVP also exists which is more adapted to describers. It has the same options as above, but without interaction during playback. A simple command line reads a configuration file that specifies the options and performs a static dubbing of the desired layers directly into the video and makes it ready for playback using any player.

5 Conclusion

This paper has reported on the ongoing development of computer-vision software tools to assist a describer in the production of sound tracks that describe the visual content of a film for people with vision loss. At this point, the first phase of the project has been completed, and a proof of concept has been implemented that automatically processes a film offline, extracts visual information such as shot transitions, key-texts, key-places, key-faces and the action of actors, assigns a text to each visual element added to the audio band, and optionally reads it using a voice synthesizer. Although the outcome is not yet completely

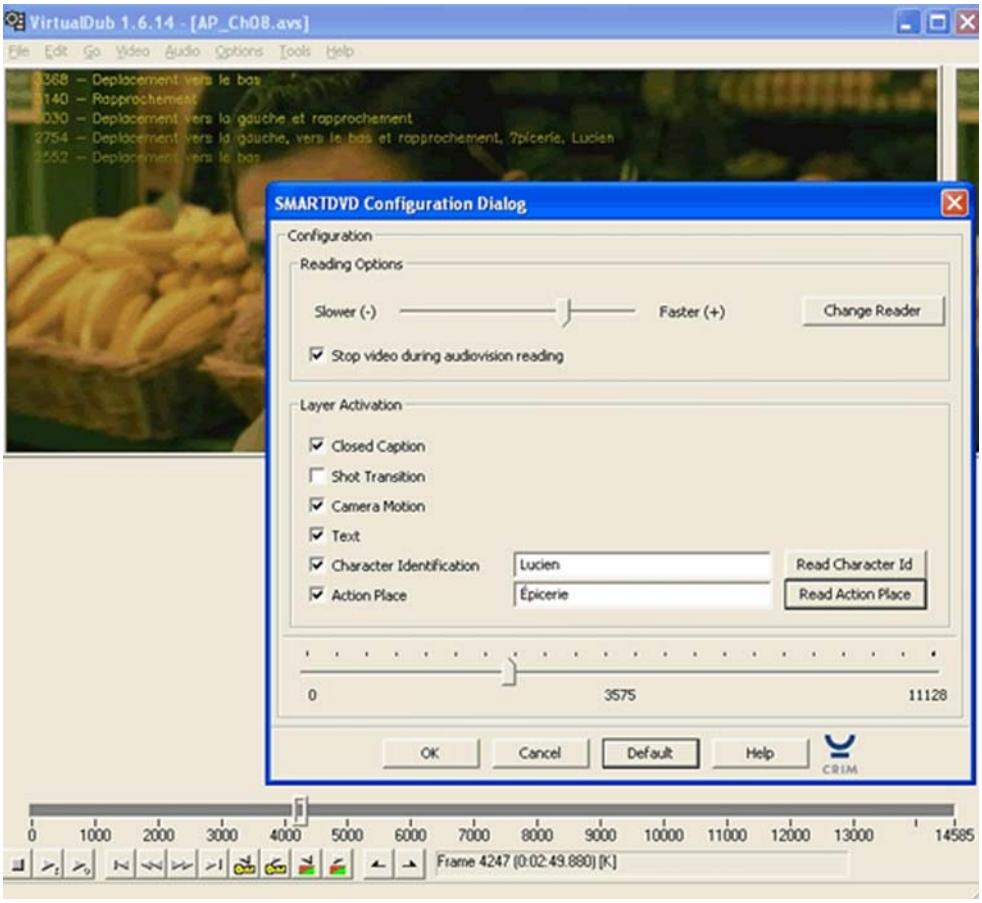

Fig. 14 Configuration window of the dynamic version of the audio-vision player. Various types of description can be activated, as well as the reading speed

satisfactory, nevertheless the technical feasibility of the entire process has been clearly demonstrated.

User consultations and production analysis have provided the backbone of the project. The objective was to gain knowledge about what types of information are really useful to this audience, in order to create guidelines for the project and for the producers of video description. This allowed identifying the important information to encode, which can be essentially summarized in four words: who, what, when and where. It is important to present the principal actors (key-faces) as early as possible after the film begins, in order to help set in the scene. This is also the case with key-places. The more often a particular location appears in the movie, the more this key-place is likely to offer meaningful details that will enable LDA to detect its contents as a predominant topic. This process is similar to what humans do when they assimilate key-places in a movie. If a location is shown several times and has a lot of specific details, the viewer quickly considers it to be a reference location.

A useful feature of the adopted key-place detection approach is that it extracts semantic characteristics. In addition to detection of places, it could also be used to detect and describe automatically specific objects in images. This will certainly be explored further, along with many other aspects such as detection of face profiles, tracking moving text, segmenting music, interpreting facial expressions, analyzing complex shot transitions, using voice commands with the video description player, etc.

The results obtained for action detection and characterization are preliminary. There are still many issues to address, including detection of the action zoom of the camera, robustness of the clustering, and labeling different groups as a single object. Using the information from this process, future work will involve trying to extract, analyze and synthesize the descriptions for the movements of the actors present in a shot, in a scene, and in the entire movie.

Thus far the feedback received from video description producers concerning the developed prototype is encouraging. The second phase of the project has now started, which aims a beta version for 2009, with improved robustness and new visual content extraction functionalities.

Acknowledgments This work is supported in part by the Department of Canadian Heritage (http://www.pch.gc.ca) through Canadian Culture Online, and the Ministère du développement économique, de l'innovation et de l'exportation (MDEIE) of the Gouvernement du

Québec (http://www.mdeie.gouv.qc.ca). The authors are very grateful to the reviewers for their constructive comments, which helped improve the quality of the paper.

References

1. Canadian Radio-television and Telecommunications Communication: Broadcasting Decision CRTC 2002-384. http://www.crtc.gc.ca/archive/ENG/Decisions/2002/db2002-384.htm (2002)
2. Piety, P.J.: The language system of audio description: an investigation as a discursive process. J. Vis. Impair. Blind. **98**(8), 1–36 (2004)
3. Turner, J.M.: Some characteristics of audio description and the corresponding moving image. In: Preston, C.M., Medford, N.J. (eds.) Proceedings of the 61st ASIS Annual Meeting, Pittsburgh, 24–29 October 1998, Information Today, pp. 108–117 (1998)
4. Turner, J.M., Colinet, E.: Using audio description for indexing moving images. Knowl. Org. **31**(4), 222–230 (2004)
5. Office of Communication: ITC guidance on standards for audio description. http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/audio_description/index.asp.html (2000)
6. Canadian Network for Inclusive Cultural Exchange: Online video description guidelines. http://cnice.utoronto.ca/guidelines/video.php (2005)
7. Guidelines for video description. http://www.joeclark.org/access/description/ad-principles.html
8. Mathieu, S.: Audiovision Interactive et Adaptable, Technical Report for the E-inclusion Research Network (2007)
9. Gagnon, L., Foucher, S., Laliberté, F., Lalonde, M., Beaulieu, M.: Towards an application of content-based video indexing to computer-assisted descriptive video. In: Proceedings of Computer and Robot Vision 2006, 8 pp (on CD-ROM) (2006)
10. Héritier, M., Gagnon, L., Foucher, S.: Places clustering of full-length film key-frames using latent aspects modeling over SIFT matches. IEEE Trans. Circuits Syst. Video Technol. (to appear) (2008)
11. Foucher, S., Gagnon, L.: Automatic detection and clustering of actor faces based on spectral clustering techniques. In: Proceedings of Computer and Robot Vision 2007, 8 pp (on CD-ROM) (2007)
12. Lalonde, M., Gagnon, L.: Key-text spotting in documentary videos using Adaboost. In: Proceedings of the IS&T/SPIE Symposium on Electronic Imaging: Applications of Neural Networks and Machine Learning in Image Processing X (SPIE #6064B) (2006)
13. Branje, C., Marshall, S., Tyndall, A., Fels, D.I.: LiveDescribe. In: Proceedings of the AMCIS 2006 (2006)
14. TRECVID. http://www-nlpir.nist.gov/projects/trecvid/
15. State-of-the-art on Multimedia Search Engines, Technical Report D2.1. Chorus Project Consortium (2007)
16. CIMWOS project. http://www.xanthi.ilsp.gr/cimwos
17. SCHEMA network of excellence. http://www.iti.gr/SCHEMA/index.html
18. VIZIR project. http://vizir.ims.tuwien.ac.at/index.html
19. Center for Digital Video Processing. http://www.cdvp.dcu.i.e
20. CALIPH and EMIR project. http://caliph-emir.sourceforge.net
21. IBM VideoAnnEx project. http://www.research.ibm.com/VideoAnnEx
22. Ricoh MovieTool project. http://www.ricoh.co.jp/src/multimedia/MovieTool
23. IBM Marvel project. http://mp7.watson.ibm.com/marvel
24. MADIS project. http://madis.crim.ca
25. Gagnon, L., Foucher, S., Gouaillier, V., Brousseau, J., Boulianne, G., Osterrath, F., Chapdelaine, C., Brun, C., Dutrisac, J., St-Onge, F., Champagne, B., Lu, X.: 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)
26. Foucher, S., Héritier, M., Lalonde, M., Byrns, D., Chapdelaine, C., Gagnon, L.: Proof-of-concept software tools for video content extraction applied to computer-assisted descriptive video, and results of consultations with producers, technical report, CRIM-07/04-07, 2007 (2007)
27. Mathieu, S., Turner, J.M.: Audiovision interactive et adaptable, technical report, 2007. http://hdl.handle.net/1866/1307 (2007)
28. Turner, J.M., Mathieu, S.: Audio description for indexing films, World Library and Information Congress (IFLA), Durban. http://members.e-inclusion.crim.ca/files/articles/IFLA-en.pdf (2007)
29. Fels, D.I., Udo, J.P., Diamond, J.E., Diamond, J.I.: A first person narrative approach to video description for animated comedy. J. Vis. Impair. Blind. **100**(5), 295–305 (2006)
30. Vendrig, J., Worring, M.: Systematic evaluation of logical story unit segmentation. IEEE Trans. Multimed. **4**(4), 492–499 (2000)
31. Bovik, A.C. (ed.): Handbook of Image and Video Processing. Academic Press, New York (2000)
32. Schaffalitzky, F., Zisserman, A.: Automated location matching in movies. Comput. Vis. Image Underst. 42:236–264 (2003)
33. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: SIGIR (1999)
34. Bosch, A., Zisserman, A., Munoz, S.: Scene Classification via pLSA. In: ECCV (2006)
35. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Van Gool, L.: Modeling Scenes with Local Descriptors and Latent Aspects. In: ICCV (2005)
36. Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: CVPR (2005)
37. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects categories in image collection, MIT AI Lab Memo AIM-2005-005 (2005)
38. Lowe, D.G.: Distinctive Image Features from Scale-invariant Keypoints. In: IJCV (2004)
39. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
40. Ng, A.Y., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. In: NIPS (2002)
41. Viola, P., Jones, M.: Robust real-time face detection. IJCV **57**(2) (2004)
42. Gagnon, L., Laliberté, F., Foucher, S., Laurendeau, D., Branzan Albu, A.: A System for Tracking and Recognizing Pedestrian Faces using a Network of Loosely Coupled Cameras, SPIE Defense and Security: Visual Information Processing XV (SPIE #6246), Orlando (2006)
43. Yang, J., Zhang, D., Frangi, A.F., Yanf, J.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. Trans. Pattern Anal. Mach. Intell. **26**(1), 131–137 (2004)
44. Kong, H., Li, X., Wang, L., Teoh, E.K., Wang, J.G., Venkateswarlu, R.: Generalized 2D principal component analysis. In: IEEE International Joint Conference on Neural Networks (IJCNN) (2005)
45. Zhang, D., Zhou, Z.H., Chen, S.: Diagonal principal component analysis for face recognition. Pattern Recognit. **39**(1), 140–142 (2006)
46. Sato, T., Kanade, T., Hughes, E.K., Smith, M.A., Satoh, S.: VideoOCR: indexing digital news libraries by recognition of superimposed caption. ACM J. Multime. Syst. **7**(5), 385–395 (1999)
47. Lienhart, R., Wernicke, A.: Localizing and segmenting text in images and videos. IEEE Trans. Circuits Syst. Video Technol. **12**(4), 256–268 (2002)
48. Chen, X., Yuille, A.L.: Detecting and Reading Text in Natural Scenes. In: CVPR, vol. II, pp. 366–373 (2004)

49. http://www.up.univ-mrs.fr/veronis/data/bigrammes.html
50. Ouellet, D., Nguyen, N.T., Dung, V.V., Laurendeau, D.: Gait and Gesture Description, Technical Report, Laval University (2007)
51. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
52. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features, Carnegie Mellon University Technical Report CMU-CS-91-132 (1991)
53. Birchfield, S.: KLT: an Implementation of the Kanade-Lucas-Tomasi Feature Tracker. http://www.ces.clemson.edu/~stb/klt
54. Bailer, W., Schallauer, P., Thallinger, G.: Camera Motion Detection, Joanneum Research. In: TRECVID (2005)
55. Birchfield, S.: Derivation of Kanade-Lucas-Tomasi Tracking Equation. http://www.ces.clemson.edu/~stb/klt/birchfield-klt-derivation.pdf (unpublished) (1997)
56. Bezdec, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York (1981)
57. Rote, G.: Computing the minimum Hausdorff distance between two point sets on a line under translation. Inf. Process. Lett. **38**, 123–127 (1991)
58. AVISynth. http://avisynth.org