Designing Caption Production Rules Based on Face, Text and Motion Detections

C. Chapdelaine*, M. Beaulieu, L. Gagnon
R&D Department, Computer Research Institute of Montreal (CRIM),
550 Sherbrooke West, Suite 100, Montreal, QC, CANADA, H3A 1B9

ABSTRACT

Producing off-line captions for the deaf and hearing impaired people is a labor-intensive task that can require up to 18 hours of production per hour of film. Captions are placed manually close to the region of interest but it must avoid masking human faces, texts or any moving objects that might be relevant to the story flow. Our goal is to use image processing techniques to reduce the off-line caption production process by automatically placing the captions on the proper consecutive frames. We implemented a computer-assisted captioning software tool which integrates detection of faces, texts and visual motion regions. The near frontal faces are detected using a cascade of weak classifier and tracked through a particle filter. Then, frames are scanned to perform text spotting and build a region map suitable for text recognition. Finally, motion mapping is based on the Lukas-Kanade optical flow algorithm and provides MPEG-7 motion descriptors. The combined detected items are then fed to a rule-based algorithm to determine the best captions localization for the related sequences of frames. This paper focuses on the defined rules to assist the human captioners and the results of a user evaluation for this approach.

Keywords: E-accessibility, Eye-tracking, TV captioning, Image processing, Video object detection.

1. INTRODUCTION

This paper presents the architecture of an implemented prototype (called SmartCaptioning) developed to assist human during off-line caption production. It demonstrates how the interactions between automatic image processing algorithms for video object detections (shot transitions, faces and texts) and motion mapping are used to identify potential captioning areas. It also explains how a production rule engine (PRE) fires sets of localization rules to automatically position caption on the image.

Deaf and hearing impaired people rely on caption to be informed and enjoy television. Skilled people (captioners) produce caption by transcribing what is being heard and by positioning the text for efficient reading without masking any visual information. Caption can be produced either on-line during the broadcast or off-line.

Captioners edit caption to establish accuracy, clarity and proper reading rate. They also have to place caption based on their assessment of the value of the visual information. Typically, they place the caption not to far from the source of speech (i.e. the speakers) and try not masking any visual element that may be relevant to the understanding of the content. Therefore, this task can be quite labor-intensive; it could require up to 18 hours to off-line caption per hour of content. The complete process offers a higher presentation quality than on-line caption which is not edited and positioned.

Our goal is to reduce the production time by automatically suggesting a position for the captions on a number of consecutive frames using a set of production rules based on image processing (IP) techniques. Furthermore, when the IP detections can be processed in real-time, these production rules could also be applied to on-line caption and thus, upgrade the presentation quality as well.

The paper is organized as follows. Section 2 describes off-line captions presentation styles and its positioning. Section 3 explains the validation of automatic detection region of interest (ROI) by conducting an eye-tracking study. Section 4 details on the implementation architecture and the PRE. Finally, we conclude by discussing our future work based on producers' feedback on SmartCaptioning prototype.

* claude.chapdelaine @crim.ca; phone 514 840 1234; fax 514 840 1244; www.crim.ca/vision

2. OFF-LINE CAPTION

Time required to produce off-line caption varies depending on the complexity of the subject, the speaking rate, the number of speakers and the rate and length of the shot. Trained captioners prepare transcripts that are split off into smaller text units to create a caption line of varying length depending on the defined presentation style. In Canada, these styles are described in guidelines [1] and distributed by The Canadian Association of Broadcasters (CAB). For off-line caption, two styles are recommended: the pop-up and the roll-up.

In a pop-up style, caption appears all at once on one to three lines. Each caption instance has to be placed on a series of consecutive frames. They can have various layouts and can appear anywhere on the image. Furthermore, the location has to change for each pop-up so that the readers can perceive the change. All of which create large production constraints on the captioners.

In a roll-up style, text units appear one line at the time over two or three lines located over a static region. In this style, the last line pushes the first line up and out. The roll-up movement indicates the changes in caption line.

CAB's guidelines contain certain recommendations on the localization of caption on the image. It is up to the producers to develop the expertise to efficiently place caption to avoid masking visual ROI. The recommendations adopted in our SmartCaptioning system are:

- Off-line caption, whether they are presented on pop-up or roll-up style, should begin and end with shot changes whenever possible, i.e. caption should appear on the first frame of the shot and could be prior to start of actual speech.
- The last caption of a shot should extend to the first frame of the next shot.
- Captions could overlap on multiple short shots if they are related to the same speaker.
- Pop-up caption of 32 characters should be visible for minimum of 1.5 seconds.
- No caption should be displayed less than one second.
- Roll-up caption of 32 characters should be visible for a minimum of 1 second.
- Roll-up caption less than 32 characters should be visible for at least 20 frames.
- The last caption of the last shot should remain visible for two seconds.
- No caption should be visible for longer than three seconds.

3. VALIDATION OF THE ROI

The expertise of efficiently positioning caption is largely based on human assessment of what is relevant to the understanding of the visual content. Consequently, to adequately assist captioners in their task, we need to ascertain that the detectable elements obtained from the various image processing techniques were indeed the same relevant ROI for the viewers. This was done trough a previous eye-tracking study involving hearing and hearing impaired people [2].

3.1 Eye-tracking

Eye-tracking analysis is one of the research tools that enable the study of eye movements and visual attention. It is known that humans set their visual attention to a restricted number of areas in an image [3] [4] [5]. Even when viewing time is increased, focus remains on those areas which are most often highly correlated among viewers. Furthermore, works in intelligent image processing shown that it is possible to define computational models to detect visual ROI similar to those selected by a human [6]. Eye-tracking analysis was revealed to be a valid method to compare the outputs of automatic image detections to human ROI [7]. In a specific visual context, such as reading or watching television, the group of ROIs representing each context defines the viewers' visual attention strategy.

We conducted an eye-tracking analysis involving 18 participants (nine hearing and nine hearing-impaired) who viewed a dataset of captioned videos representing five types of television content. The results of this work are available in [2]. The conclusions integrated in our production rules are:

- Faces should always be identified as a ROI even in the case of multiple faces on the images. Caption should not hide faces and should be placed close enough to facilitate viewing especially for the hearing impaired viewers
- The detection algorithms need to include moving objects, especially for sports.
- Close-up and blurred objects should not be considered as ROI. This implies that detection of camera movement should be integrated in the motion mapping.

Eye-tracking analysis validated that the IP detections could match human ROI since it confirmed a large number of detectable ROI and it also helped identifying ROI as well.

4. SMART CAPTIONING ARCHITECTURE

The SmartCaptioning system (Fig. 1) is designed to take care of the repetitive and procedural work during the captioning process, leaving the creativity and decision tasks to the captioner.

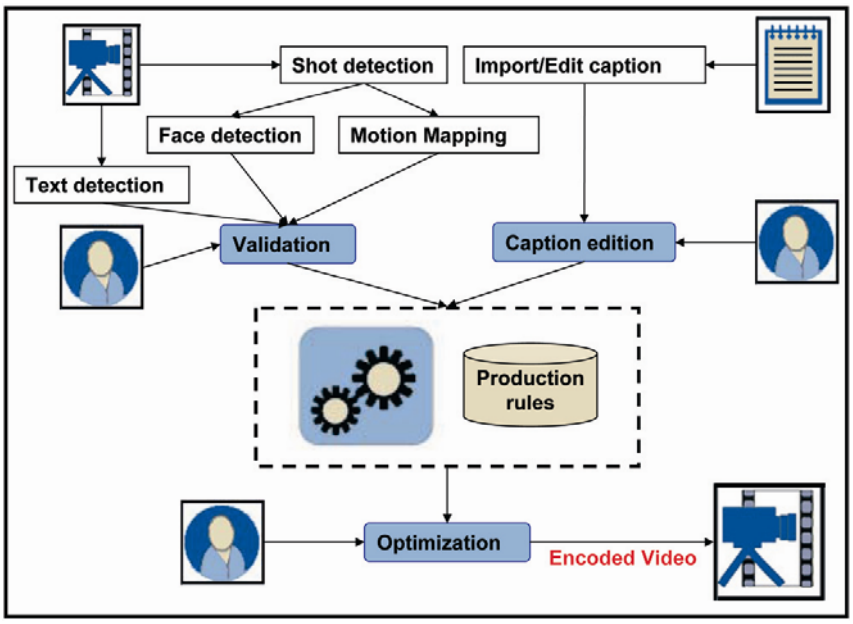

Fig. 1. SmartCaptioning system architecture

In order to be able to properly assist captioners in their work, the Smart Captioning system architecture integrates different image processing techniques such as shot transition, face and text detection as well as motion mapping. When the detections and mapping are done, the captioner validates the data keeping only the relevant items from the detection. Once the validation is completed, the captioner triggers the rule-based process to generate an automated caption version. The produced version can be viewed by the captioner and edited if necessary. Indeed, in cases where the system cannot apply a rule satisfying all the existing conditions or concurrent rules, a default position is provided to the captioner for validation. The next sub-sections give an overview of the main parts of the system.

4.1 Shot detection

Detection of video shot transitions is a necessary step in most video analysis systems. It aims at finding video segments having a homogeneous visual content. We automatically detect shot transitions based on the mutual color information between successive frames, calculated for each RGB components [8]. Cuts are identified if intensity or color is abruptly changed. The detection of gradual transitions is much more complex and most detectors have poor performances. In our system [9], we focus on cuts to synchronize caption and to serve as an input for other tasks to trigger or reset a particular algorithm.

Shot detection is used first in the planning process, to get a sense of the content rhythm to be processed. Many short consecutive shots indicate many synchronization and short delays, thus implying a more complex production. Second, it is used in the production rules, to associate captions and shot. Each caption is associated to a shot and the first one is synchronized to the beginning of the shot even if the corresponding dialogue comes later in the shot. Also the last caption is synchronized with the last frame of a shot. Finally, shot detection is also used in the system process, to temporally segment the video. Shots are the film segments used by the other detection techniques. Thus, shot detection is done first and serves as an input to all the others processes.

4.2 Face detection

Face detection is done on near-frontal views defined by a cascade of weak classifiers [10] [11]. Face tracking is done through a particle filter and generate trajectories (Fig. 2). As proposed in [12], the particle weight for a given ROI depends on the face classifier response. For a given ROI, we take the classifier response as the maximum level reached in the weak classifier cascade (the maximum being 24). Details of the face detection and tracking implementation can be found in [9]. For each shot, we keep the first frame of each trajectory.

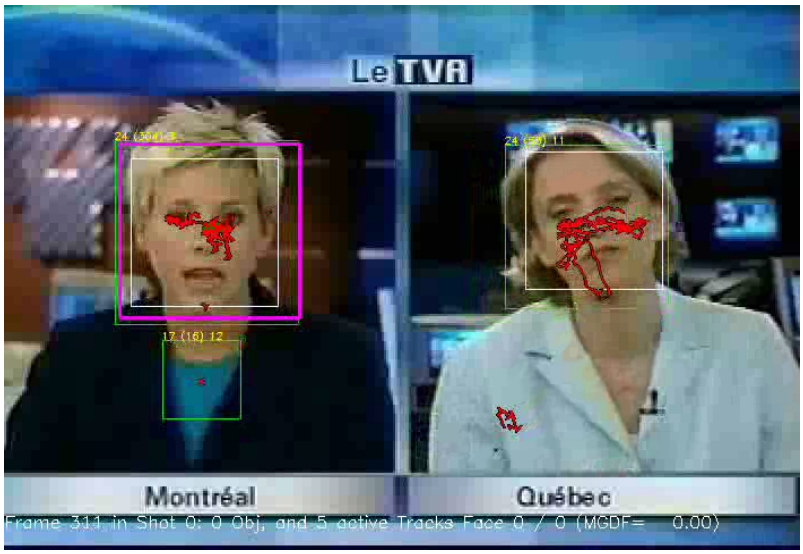

Fig. 2. Face detection (in square boxes) and tracking (lines in red).

In our system, face detection is used first in planning to indicate the number of people present in the shot and within the whole content; second in the production rules, to associate caption to a particular person and finally in the system process, to reduce the region for text and motion detection.

4.3 Text detection

Text detection is based on a cascade of classifiers trained with Adaboost [13]. Simple features (e.g. mean/variance ratio of grayscale values and x/y derivatives) are measured for various sub-areas upon which a decision is made on the presence/absence of text. The result for each frame is a set of text ROI where text is expected to be found (Fig. 3). The ROI are pre-processed before OCR to remove their background and noise. Our strategy for segmenting into sub-areas is

to consider the centroid pixels of any detected ROI that participated in the aggregation step of the text detection stage. The RGB values of these pixels are then collected into a set associated to their sub-area. A K-means clustering algorithm is invoked to find the three dominant colors (foreground, background and noise). Then, character recognition is performed by commercial OCR software. In the same manner as in the face detection, the text bounding boxes detected are than used as input into the motion detection to reduce the amount of computation.

Fig. 3. Text detection and OCR

4.4 Motion mapping

Usually, captions should avoid regions with high motion activity because they might indicate important information to the viewer. Our approach consists then in producing a map where captions can be displayed.

The motion detection algorithm is based on the Lukas-Kanade optical flow technique [14]. Optical flow is computed between two frames at each pixel to obtain its mean velocity magnitude as well as others various global and local motion information (Fig. 4).

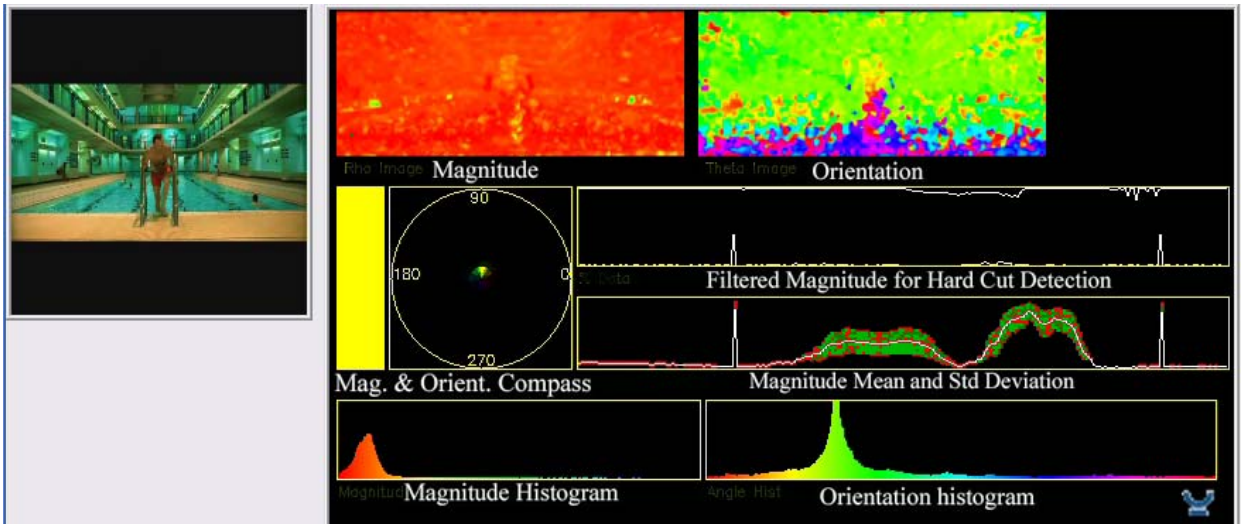

Fig. 4. Example of motion detection information provided.

The optical flow results are used to perform foreground detection and mask regions where no movement is detected between two frames (Fig. 5), thus generating a Motion Activity Map (MAM) for each frame.

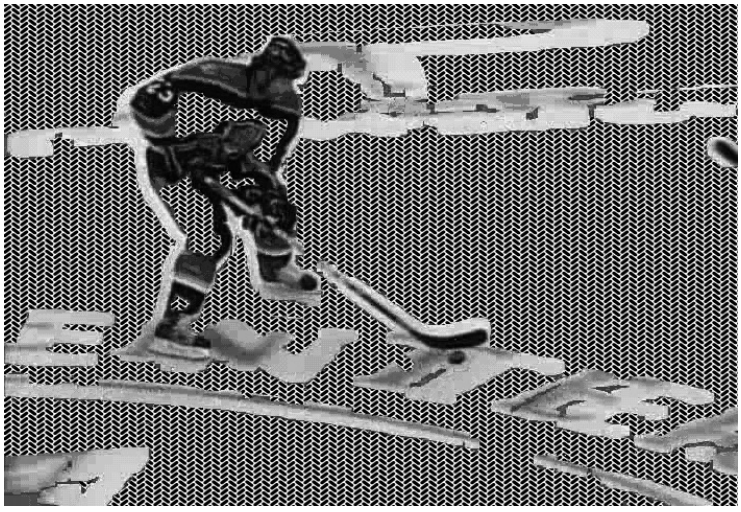

Fig. 5. Example of MAM for a frame

4.5 Motion descriptors

The optical flow results are also used to identify critical frames in a shot, i.e. those with the highest velocity magnitude. The critical frames are used to build a Motion Activity Grid (MAG) which partitions the frame into sub-sections where caption could potentially be placed. For each frame, 64 sub-sections are defined based on the television format and usage [15] of the Society of Motion Picture and Television Engineers (SMPTE). These sub-sections are included in the "save title area" (STA) portion of the production aperture defined in the SMPTE.

For example, for a digital format of 720x486 pixels, the STA would be of 576x384 pixels. Giving that a caption line has a height of 24 pixels, this defines a MAG of 16 potential lines. The number of columns has to support the maximum of 32 characters per caption line which is grouped into regions large enough to place few words (4 groups of 144 pixels). So, the MAG of each frame is a 16x4 grid, totalizing 64 areas of magnitude velocity mean and direction (Fig. 6).

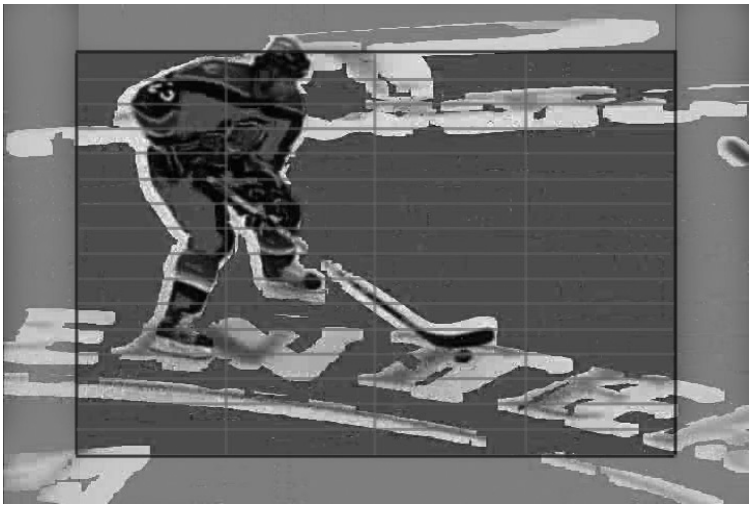

Fig. 6. Example of MAG of a critical frame

4.6 Human interaction

Once all the IP detections are done, the system informs the captioner and displays the results for validation and identification. Captioner has to reject the false alarms or others instances of detection that they may judge irrelevant. For example, a face in a picture frame on a wall could be detected but not be associated with any caption.

5. PRODUCTION RULES

Three sets of production rules are formed according to the caption position in the shot: 1) beginning of the shot, 2) during the shot and 3) end of shot. The first step is to identify the visible sequence of frames (VSF) which is the number of frames during which a particular caption will be visible (Fig. 7). The VSF varies according to the number of characters and respect the standards from the CAB's guidelines. Second, a first estimation of the caption region (CR) is determined based on the face and text detections results of the VSF.

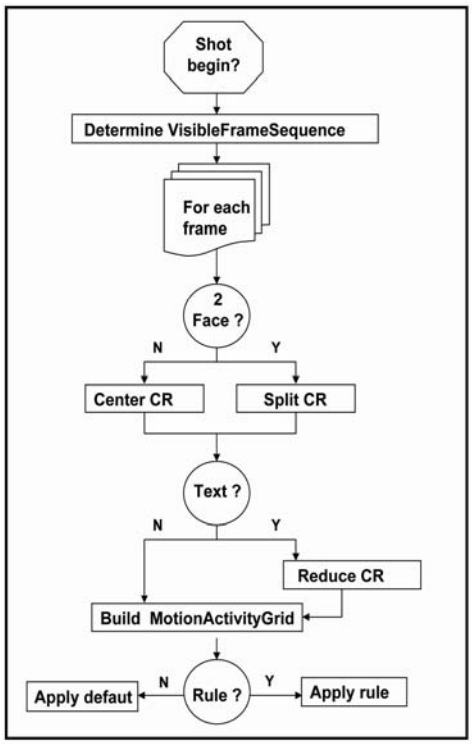

Fig. 7. Example of a rule application

Finally, the MAG of each critical frame in the VSF is examined to further identify the proper caption region. As shown in Fig. 8, a caption that has to been placed between frame 63 and 170, would build a MAG starting at the frame having the highest activity (113 here) to ensure that caption would not mask important motion in the shot.

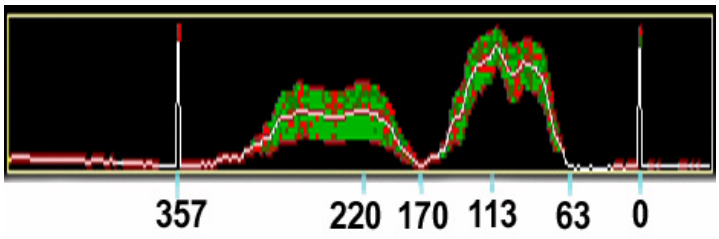

Fig. 8. Example of motion activity real-time feedback

6. CONCLUSION

As presented in this paper, building a production caption system requires implementing rules of production not only based on existing guidelines but also from inputs from the intended users. Indeed, our eye-tracking analysis validated the integration of IP detections by confirming a larger number of detectable ROI and also by identifying others as well. The guidelines and the results of detections enable the development of a caption production engine that succeeded in producing captioned data with less human intervention. Still, the implementation and the evaluation of the SmartCaptioning system provide us with meaningful specifications for the continuation of our work.

Future works involve making shot detection more robust to gradual transition, reducing false alarms for face and text detection, improving motion detection and adding more specialized production rules. Furthermore, we intend to produce more contents with the collaboration of our partners in order to undergo evaluations on the production process with producers and of the produced content with the potential viewers.

ACKNOWLEDGEMENTS

This work is supported in part by (1) the Department of Canadian Heritage (www.pch.gc.ca) through the Canadian Culture Online program and (2) the Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE) du Gouvernement du Québec. We would also like to give our sincere thanks to all the hearing impaired and hearing individuals who gave us their time for their patience and their kind participation.

REFERENCES

[1] Canadian Association of Broadcasters, *Closed Captioning Standards and Protocol*, CAB eds., 2004.

[2] C. Chapdelaine, V. Gouaillier, M. Beaulieu, L. Gagnon, "Improving Video Captioning for Deaf and Hearing-impaired People Based on Eye Movement and Attention Overload", Proc. of SPIE, Volume 6492, Human Vision & Electronic Imaging XII, (2007).

[3] A. L. Yarbus. *Eye Movements and Vision*, Plenum Press, New York NY, 1967.

[4] M. I. Posner and S.E. Petersen, "The attention system of the human brain (review) ", Ann. Rev. Neurosciences, **(13)**, 25-42. (1990).

[5] J. Senders. "Distribution of attention in static and dynamic scenes", In Proceedings SPIE 3016, pages 186-194, San Jose (1997).

[6] C. M. Privitera, L. W. Stark, "Algorithms for defining visual regions-of interest: Comparison with eye fixations", IEEE Trans. Pattern Analysis and Machine Intelligence, **vol. 22, no. 9**, pp. 970-982, (2000).

[7] R. B. Goldstein, R. L. Woods, E. Peli, "Where people look when watching movies: Do all viewers look at the same place?", Computers in Biology and Medicine. Published online 28 September (2006).

[8] Z. Cerneková, I. Pitas, C. Nikou, "Information Theory-Based Shot Cut/Fade Detection and Video Summarization", IEEE Trans. On Circuits and Systems for Video Technology, V**ol. 16, No. 1**, pp. 82-91, (2006).

[9] S. Foucher, L. Gagnon, "Automatic Detection and Clustering of Actor Faces based on Spectral Clustering Techniques", CRV, pp. 113-120, (2007).

[10] P. Viola, M. J. Jones, "Rapid object detection using a boosted cascade of simple features", CVPR, pp. 511-518, (2001).

[11] E. Lienhart, J. Maydt, "An extended Set of Haar-like Features for Rapid Object Detection", ICME, (2002).

[12] R. C. Verma, C. Schmid, K. Mikolajczyk, "Face Detection and Tracking in a Video by Propagating Detection Probabilities", IEEE Trans. on PAMI, **Vol. 25, No. 10**, (2003).

[13] M. Lalonde, 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), (2006).

[14] B. Lucas, T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", Proc. of 7th International Joint Conference on Artificial Intelligence, pp. 674-679, (1981).

[15] Society of Motion Picture and Television Engineers, http://www.smpte.org/home.