2012 IEEE International Conference on Multimedia and Expo

A local temporal context-based approach for TV news story segmentation

Émilie Dumont and Georges Quénot
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
Email: Emilie.Dumont@imag.fr - Georges.Quenot@imag.fr

***Abstract*—Users are often interested in retrieving only a particular passage on a topic of interest to them. It is therefore necessary to split videos into shorter segments corresponding to appropriate retrieval units. We propose here a method based on a local temporal context for the segmentation of TV news videos into stories. First, we extract multiple descriptors which are complementary and give good insights about story boundaries. Once extracted, these descriptors are expanded with a local temporal context and combined by an early fusion process. The story boundaries are then predicted using machine learning techniques.**
We investigate the system by experiments conducted using TRECVID 2003 data and protocol of the story boundary detection task and we show that the extension of multimodal descriptors by a local temporal context approach improves results and our method outperforms the state of the art.

***Keywords*-news video segmentation, story detection, local temporal context, descriptor extraction, machine learning techniques**

I. Introduction

With the development of digital technology and extensive applications, video has become an important way of delivering information. However, finding a video content corresponding to a particular user's need is not always easy for a variety of reasons, including poor or incomplete content indexing. Also, while video content is often stored in rather large files or broadcasted in continuous streams, users are often interested in retrieving only a particular passage on a topic of interest to them; when a user asks the system for information relevant to a query, it is insufficient for it to respond by simply pointing the user at an entire video. One would expect the system to return a reasonably short section, preferably a section only as long as necessary to provide the requested topic. It is therefore necessary to split video documents or streams into shorter segments corresponding to appropriate retrieval units, for instance a particular scene in a movie or a particular news in a TV journal. These retrieval units can be defined hierarchically on order to potentially satisfy user needs at different levels of granularity. The retrieval units are not only relevant as search result units but also as units for content-based indexing and for further increasing the content-based video retrieval (CVBR) systems' effectiveness.

A video can be analyzed at different levels of granularity. For the image track, the lower level is the individual frame which is generally used for extracting static visual descriptors like color, texture, shape, or interest points. Videos can also be decomposed into shots; a shot is a basic video unit showing a sequence of frames captured by a single camera in a single continuous action in time and space. The shot however, is not a good retrieval unit as it usually lasts only a few seconds. High-level techniques are therefore required to determine a more descriptive segment. We focus in this work on the automatic segmentation of TV journals into individual news or commercial sections if some are present. More specifically, we aim at detecting boundaries between news stories or between a news story and a commercial section. Though this work is conducted in a particular context, it is likely that it could be applied in other ones, *e.g.* the segmentation of movies, talk shows or sports events. Story segmentation allows better navigation within a video. It can also be used as the starting point for other applications such as video summaries or story search system.

We propose an approach based on multimodal descriptor extraction expanded with a local temporal context. We use the temporal information provided by local context of a video sequence to improve performance. Our system is based on the complementarities of visual and audio information from a video. The story boundary detection is generally more efficient when several and varied descriptors are used. We use machine learning methods to perform the story boundaries detection from these multiple descriptors expanded by a local temporal context.

The paper is organized as follows: we first present related works and, when possible, recall the performances of these systems. We then explain our system based on the fusion of multiple descriptors for the news story segmentation. The multimodal descriptors used are briefly presented in section IV. The local temporal context extension is detailed in section V. The early fusion of these descriptors is then presented. Finally, we present and analyze experimental results and we conclude.

II. Related works

Related works and existing solutions are developed in most cases for broadcast TV and more precisely for broadcast news. It was the case for the task proposed by TRECVID in 2003 and 2004 "Story segmentation" [1], [2] and the more recent ARGOS campaign [3]. Existing techniques for structuring a TV broadcast [4] are classified into three categories: manual approach by skilled-workers,

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.3

metadata-based approach and content-based approach. We focus on the last category. The approach we explored is to segment on the story level; the video segmentation consists in automatically and accurately determining the boundaries (*i.e.* the start and the end) of each story.

The authors of [5] presented one of the first works on video segmentation in scenes. Their point of view for scene segmentation is: first locate each camera shot and second combine shots based on content to obtain the start and end points of each scene. They focus on low level audio properties.

The method proposed by Chaisorn and al. [6], [7] obtained one of the best results at the TRECVID 2003 story boundary detection task, as they achieved a F1 measure accuracy over 0.75. They, first, segmented the input video into shots. Then, they extracted a suitable set of descriptors to model the contents of shots. They employed a learning-based approach to classify the shots into the set of pre-defined categories. Finally, they identified story boundaries using a HMM model or inductive rules. However, they selected 13 categories of shots, like Anchor, Sports, Weather, Program logo, Finance, Speech/Interview ... Although effective, their technique requires a lot of manually annotated data. The method proposed here needs much less annotated data.

Recently, authors of [8] segmented videos into stories by detecting anchor person into shots; the text stream is also segmented into stories using a Latent Dirichlet Allocation (LDA) based approach. They obtained a F1 measure equals to 0.58 on the TRECVID 2003 story boundary detection task. In the paper [9], they presented a scheme for semantic story segmentation based on anchor person detection. The proposed model uses a split and merge mechanism for finding story boundaries. The approach is based on visual descriptors and text transcripts. The performance of this method is over 0.6 for F1 measure also on the TRECVID 2003 story boundary detection task. In the study [10], a set of key events are first detected from multimedia signal sources, including a large scale concept ontology for images, text generated from automatic speech recognition systems, descriptors extracted from audio track, and high-level video transcriptions. Then a fusion scheme is investigated using the maximum figure-of-merit learning approach. They obtained a F1 measure equals to 0.651 on the TRECVID 2003 story boundary detection task.

In this paper, we propose a more effective method than the actual state of art (experimented on the same test data). Moreover, our method used almost external annotation beyond the one available during the TRECVID campaign.

III. System Overview

The architecture of the proposed approach is showed by figure 1; it is composed of four levels: the video segmentation, the descriptor extraction, the temporal context extension and the classification. In major cases, previous works used the shot as a basic segmentation for this problem. However, in TRECVID development set, there is 94.1% story boundary near a shot boundary. This means that a system working at the shot level cannot find about 6% of the story boundaries. For example, at the end of a story, an anchorperson can appear to give a summary or a conclusion and switch to another topic. In this case, there is no shot transition between the two stories. Therefore, we decide to work with one second segments as a basic unit as this unit may be considered as a basic unit of a video [11]. This segmentation has a visual meaning: a segment of one second is just long enough to be interpreted by a human brain.

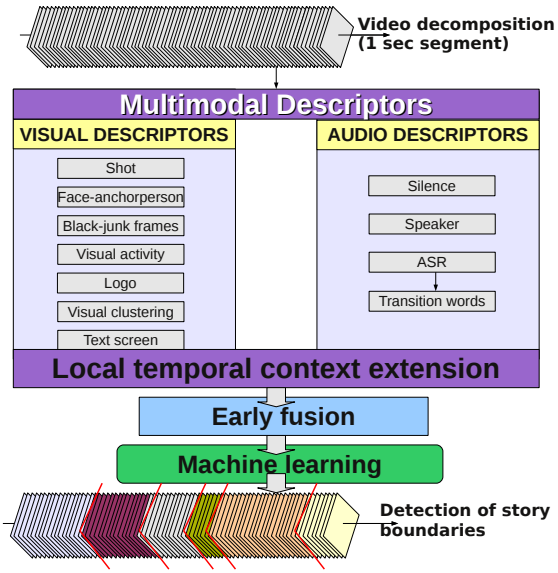

Figure 1. Multiple Descriptors Based Story Segmentation

In a second step, we extract a suitable set of multimodal descriptors to model the content of one second segments like anchorperson, logo presence, silence, transition words ... These descriptors are automatically extracted in an unsupervised way. However, most often in previous works, this step was done with methods using manual annotations.

Once extracted, these descriptors are expanded with a local temporal context. The main idea of this step is to improve the descriptors by these temporal changes. For example, the appearance or disappearance of the logo is an information more important than only the presence of the logo in the video sequence.

Finally, like in major works, we focus on finding the boundaries of every successive story in the video stream. To perform this, we test the suitability of classifiers using WEKA with default parameters to find the best one.

IV. Multimodal descriptors based News Stories Segmentation

Multimodal descriptors constitute a pool to be used for story boundary detection. They are complementary and give good insights about story boundaries. These descriptors are

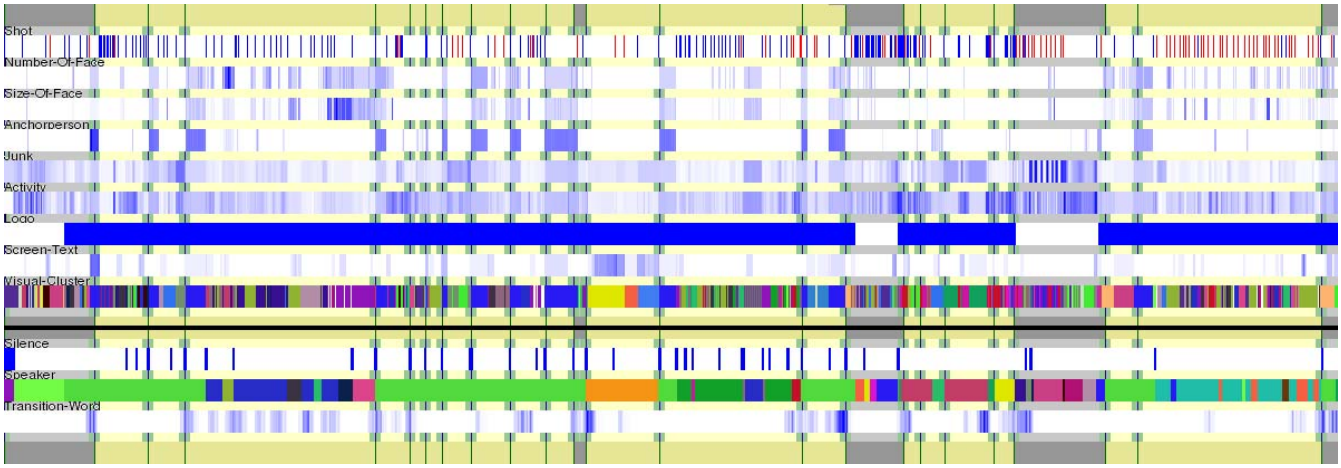

Figure 2. Example of multimodal descriptors without local temporal context extension. The top and bottom thick lines (or stripes) represent the ground truth with transitions in black and stories in light green (news) or dark gray (advertisements/misc). The similar line (or stripe) with a thick black line in the middle shows the same information while also separating the visual features (above) from the audio features (below). Thin lines between the thick ones reproduce the top and bottom thick lines but with lighter colors for the story types and additionally with a 5 second green expansion around the boundaries corresponding to the fuzziness factor associated to the evaluation metric. These are replicated so that it is easier to see how the feature values or transitions match them. Also, the beginning of the thin lines contains the name of the feature represented in the thick lines immediately below them. Finally, the remaining thick lines represent the feature values with three types of coding. For scalar analog values, the blue intensity corresponds to the real value normalized between 0 and 1. For binary values, this is the same except that only the extreme values are used and that in the case of shot boundaries, blue is used for cuts and red is used for gradual transitions. For cluster index values (clusters and speakers), a random color map is generated and used.

obtained either by a third party system like: shot boundary detection [12], face detection [13], junk and visual activity [11], screen text detection [14], silence detection, automatic speech recognition [15], speaker diarization [16]; or built for our proposal [11]. Text tiling was considered but, surprisingly, it was found that it did not help.

A. Anchorperson

In order to detect anchorperson sequences, we assume that frames with the anchor person are frames that i) contain a face centered, and ii) are very likely to appear frequently almost "as is" in the video. Consequently, we first select the frames that contain a centered face as candidates to be an anchorperson template. For a given video and in order to select an appropriate anchorperson template, we expect the average visual similarity of candidates with a pre-fixed percentage of candidates to be maximal and choose the template as the frame that exhibits the greatest similarity. Finally, the similarity between the template and a frame is used like a confidence measure of the presence of an anchorperson.

B. Logo

A TV logo is a graphic representation which is used to identify a channel. A logo is placed in the same place and continuously, except during commercials. Based on this observation, we compute the average frame of the video and the variance of the pixel color in the video. Pixels with the lowest variance are considered to be a part of the logo. Their position will be called the reference position. During the logo detection step, for a given frame, the absolute difference between the colors of the pixels situated at a reference position and their counterpart in the average image is computed. The lower the sum is, the more probable the logo is in the frame. In order to reduce the computation time, we manually selected the search region for each different channel.

C. Transition words

Based on the ASR, we extract the most frequent transition words. We first remove all stop words from the transcription. Then, we select the most frequent words that appear in a temporal window that overlaps a story transition. Finally, for each selected word w, we determine a score related to the non uniform probability to find a transition at time $t+i$ sec given that w were pronounced at time t.

Words	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$
A.B.C	0.02	0.03	0.016	0.01	0.12	0.62	0.18
News	0.03	0.16	0.15	0.04	0.29	0.33	0.06
Tonight	0.07	0.23	0.32	0.10	0.14	0.10	0.04
Today	0.18	0.30	0.46	0.02	0.00	0.01	0.02

Table I
TRANSITION WORDS AND THEIR SCORES

Table I shows results obtained on ABC videos. If i ranges between -3 and $+3$ seconds, we can notice that the extracted words are ABC, News, Today and Tonight, ABC and News being pronounced one or two seconds after

a transition while Today and Tonight appears a few seconds before a transition.

D. Multimodal descriptors

Figure 2 shows a representation of multimodal descriptors. Each pixel column corresponds to a one-second segment. The shot detection information is decomposed into two binary values: the first one represents the presence of a cut transition and the second represents the presence of a gradual transition in the one-second segment. The presence of silence and logo are represented by a binary value. Visual cluster and speaker are represented by the cluster index. And finally, other descriptors are numerical values.

As it can be seen, silence is well correlated with the ground truth although it lacks precision (it detects a silence between the first two story boundaries). This false alarm can nevertheless be corrected using other descriptors like for example anchorperson or shot transition. The combinatorial is very complex, so we rely on an automatic procedure to combine these descriptors.

V. Local temporal context extension

All descriptors are extracted for each one second segments of video. Therefore, they do not take into account the temporal information included in a video. Certainly, the information of the presence or absence of a descriptor is important, but the information about the appearance or disappearance is more relevant. Based on this observation, we extend the descriptors with a local temporal context, more precisely by the descriptor values of closest segments.

Descriptors presented in the previous section are used to build a composite vector as an input for a classifier. We propose to improve descriptors by adding a local temporal context. We use a strategy based on a sliding window: for a one second segment s coming into sight at time t in the video, we use a sliding window with a fixed length equal to $2l+1$ and where the current segment is located at the center of the window $W_s = \{s_{t-l}, ...s_t, ..., s_{t+l}\}$. For a sliding window, we extract three categories of representations:

- the list V_{all} of all values containing in the sliding window ($2l + 1$ values);
- the list V_{diff} of the difference between each couple of one second segment with an equal distance to s_t plus the central value s_t itself ($l + 1$ values);
- V_{gauss} the values of the Gaussian distribution, the derivation of Gaussian distribution and the second derivation of Gaussian distribution (3 values).

The first solution corresponds to feeding the classifier with an input vector which is a concatenation of a number of column vectors around the current one or to use a vertical slice of several columns in the representation given in figure 2. This is the most complete information that can be passed on and it leave open to the classifier underlying machine learning method to decide whether it will use for each feature either the single central value, the level around it, the variation around it or any combination of them including how far around it should go. Though this is the most complete, it is also the most costly one and not necessarily the most efficient one. As we can have the intuition that either the level, the variation or a combination of both can be more compact and more synthetic we considered the two other possibilities, the third one being even more compact and synthetic than the second one. We also considered the possibility of optimizing the size of the window and the neighborhood representation type by tuning them using a development set.

Finally, each multimodal vector used as input for the classifier is a concatenation of the best descriptors' representations, with a resulting dimension equal to 231. We chose to perform an early fusion for avoiding the loss of the correlation information between different features. We tested several classifiers using WEKA [17] for finding the most effective one.

VI. Experimental results

A. Experimental Protocol

We have experimented our approach in the context of the TRECVID 2003 Story Segmentation Task. The collection contains about 120 hours of ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998. We chose this dataset because it is the only one which is available and widely used by the community; it allows us to compare our method with the state of the art.

We developed and tuned the system only within the development set (partitioned itself into a training and a test set by a random process), then we applied it on a test set. Since story boundaries are rather abrupt changes of focus, story boundary evaluation is modeled on the evaluation of shot boundaries: to evaluate the Story Segmentation, an automatic comparison to human-annotated reference is done to extract a recall and precision measure. A story boundary is expressed as a time offset with respect to the start of the video file in seconds, accurate to nearest hundredth of a second. Each reference boundary is expanded with a fuzziness factor of five seconds in each direction, resulting in an evaluation interval of 10 seconds. If a computed boundary does not fall in the evaluation interval of a reference boundary, it is considered a false alarm.

- Story boundary recall= number of reference boundaries detected / total number of reference boundaries
- Story boundary precision= (total number of submitted boundaries minus the total amount of false alarms)/ total number of submitted boundaries
- Story boundary F1-measure= 2 $\times$ recall $\times$ precision / (recall+precision)

Descriptor	Shot	Anchor	Silence	Speaker	Face	TWord	TScreen	Junk	Activity	Logo
Length	1	21	9	15	11	13	5	9	13	21
Values	V_{all}	V_{diff}	V_{diff} V_{gauss}	V_{diff}	V_{all}	V_{all} V_{gauss}	V_{all}	V_{diff}	V_{diff}	V_{diff} V_{gauss}

Table II
BEST DESCRIPTOR REPRESENTATION. IN THIS TABLE, WE CAN SEE FOR EACH DESCRIPTOR: THE BEST LENGTH l FOR THE SLIDING WINDOW, AND THE SELECTED CATEGORIES OF VALUES.

B. Classifier selection

We made a selection of the best classifier method for our problem: 48 classifiers have been tested, for more information about these classifiers, see [18]. Figure 3 shows best results obtained in term of F1 measure on the development set. Results show that RandomForest is the best classifier

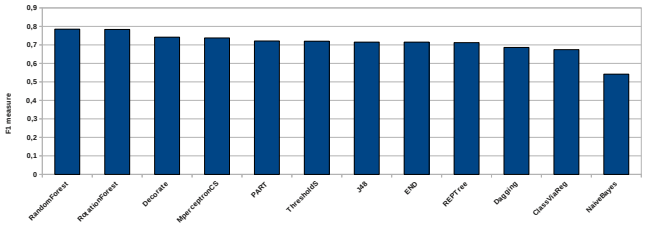

Figure 3. Results for the best classifiers

for this problem. This system is trained in 151 seconds and the predictions are computed in 32 seconds.

Results also show that the classifiers in the category of trees are on average the best in our case. This can partially be explained by the non-normalized descriptors that we used. However, this is a complex problem because our descriptors do not have the same scale. For example, it is difficult to compare the number of faces in a video segment and a confidence value of visual activity. For our problem, it is also interesting to note that the amount of positive is very low compared to the number of negative. So, classifiers like SVM are not suitable.

C. Local temporal context experiments

For each descriptor, we tested different lengths of sliding window (from 1 sec to 31 sec) and different representations (V_{all}, V_{diff} or V_{gauss}) in order to find the best combination for each descriptor (other descriptors were used without local context information). The figure 4 shows the results for three different descriptors: speaker, face and transition words. The curve "Base" represents results without local temporal extension. It is clear that the local temporal context improve the quality of the predictions. Table II shows the best combination for the selected descriptors.

In figure 5, we compare the performance of the different methods of local temporal context extractions. We can see that the local temporal context improve performance and the best results are obtained by using the best local temporal

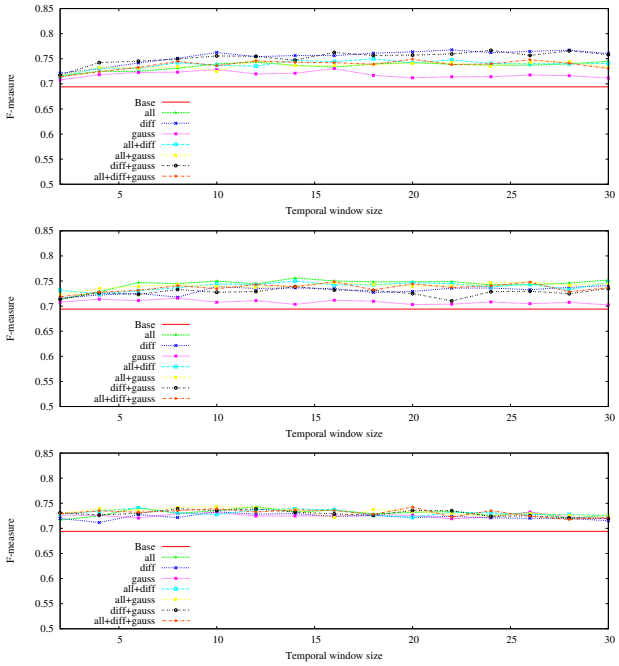

Figure 4. Results for local temporal context of a descriptor

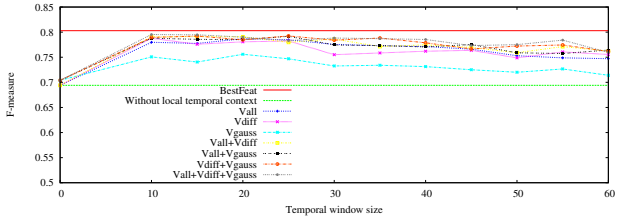

Figure 5. Results for local temporal context

context for each descriptor. This method use vectors of 231-dimensions. The closest results to this method are obtained using a sliding window equals to 15 and extracting V_{all}, V_{diff} and V_{gauss} for each descriptor; however in this case, the dimension became 650. So the selection of optimal parameter for each descriptor is more interesting.

D. Results

We compare our results with state of the art methods in table III. We have obtained a recall of 0.878 and a precision of 0.767 for the test set, which gives a F1 measure equals

to 0.819 with a threshold optimized on the development set. On the same data set, our system is more effective than the actual systems.

Method	[6]	[8]	[9]	[10]	Our Method
Recall	0.749	0.54	0.497	0.581	0.878
Precision	0.802	0.64	0.750	0.739	0.767
F1	0.775	0.58	0.600	0.651	0.819

Table III
COMPARISON WITH STATE-OF-ART

VII. CONCLUSION

We have presented a system for segmenting TV news videos into stories. This system is based on multimodal descriptors extraction. The originality of the approach is in the use of a temporal context for the descriptors before their combination by early fusion; it is also in the use of machine learning techniques for finding the candidate transitions form a large number of heterogeneous low-level descriptors.

This system has the advantage that it require no or minimal external annotation. It was evaluated in the context of the TRECVID 2003 story segmentation task and obtained better performance than the current state of the art.

Future work would include other relevant descriptors for this task and an efficient step of normalization. Descriptors of interest could be jingles, sports, weather or finance in a video collection. Regarding the method for predicting the presence of story transition, it could be improved through a process that takes into account the video structure. As a starting point, we can include in the description of a one second segment the time fraction of the video at which it occurs.

ACKNOWLEDGMENTS

This work was realized as part of the Quaero Programme funded by OSEO, French State agency for innovation.

REFERENCES

[1] A. Smeaton, W. Kraaij, and P. Over, "TRECVID - an overview," in *Proceedings of TRECVID*, 2003.

[2] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 656–659.

[3] P. Joly, J. Benois-Pineau, E. Kijak, and G. Qunot, "The ARGOS campaign: Evaluation of Video Analysis Tools," *Signal Processing: Image Communication*, vol. 22, no. 7-8, pp. 705–717, Aug. 2007.

[4] A. E. Abduraman, S.-A. Berrani, and B. Mérialdo, *TV program structuring techniques : A review*. Book chapter in TV Content Analysis: Techniques and Applications, October 2011.

[5] J. M. Gauch, S. Gauch, S. Bouix, and X. Zhu, "Real time video scene detection and classification," *Information processing and Management*, vol. 35, no. 3, pp. 381–400, May 1999.

[6] L. Chaisorn and T. seng Chua, "Story boundary detection in news video using global rule induction technique," in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 2101–2104.

[7] L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, pp. 187–208, 2003.

[8] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose, "Tv news story segmentation based on semantic coherence and content similarity," in *Proceedings of the 16th international conference on Advances in Multimedia Modeling*, 2010, pp. 347–357.

[9] A. Goyal, P. Punitha, F. Hopfgartner, and J. M. Jose, "Split and merge based story segmentation in news videos," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, 2009, pp. 766–770.

[10] C. Ma, B. Byun, I. Kim, and C.-H. Lee, "A detection-based approach to broadcast news video story segmentation," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1957–1960.

[11] E. Dumont and B. Mérialdo, "Rushes video summarization and evaluation," *Multimedia Tools Appl.*, vol. 48, no. 1, pp. 51–68, May 2010.

[12] G. Quénot, D. Moraru, and L. Besacier, "CLIPS at TRECvid: Shot boundary detection and feature detection," in *Proceedings of TRECVID*, 2003.

[13] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions On Pattern Analysis and Machine intelligence*, vol. 20, pp. 23–38, 1998.

[14] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From text detection in videos to person identification," in *IEEE International Conference on Multimedia and Expo*, 2012.

[15] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[16] V.-B. Le, O. Mella, and D. Fohr, "Speaker Diarization using Normalized Cross Likelihood Ratio," in *proceeding of INTERSPEECH 2007*, Aug. 2007, pp. 1869–1872.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, 2009.

[18] http://www.cs.waikato.ac.nz/ml/weka/.