CLIPS at TRECvid: Shot Boundary Detection and Feature Detection

Georges M. Quénot, Daniel Moraru, and Laurent Besacier

CLIPS-IMAG, BP53, 38041 Grenoble Cedex 9, France
`Georges.Quenot@imag.fr`

Abstract

This paper presents the systems used by CLIPS-IMAG to perform the Shot Boundary Detection (SBD) task and the Feature Extraction (FE) task of the TRECvid workshop. Results obtained for the 2003 evaluation are presented. The CLIPS SBD system based on image difference with motion compensation and direct dissolve detection was second among 14 systems. This system gives control of the silence to noise ratio over a wide range of values and for an equal value of noise and silence (or recall and precision), the value is 12 % for all types of transitions. Detection of person X from speaker recognition alone was deceiving due to the small number of shots containing person X in the overall test collection (about 1/700) and the even small number in which person X was actually speaking (about 1/6000). Detection of person X from speech transcription performed much better but was still lower than other systems using also the image track for the detection.

1 Introduction

The CLIPS-IMAG laboratory has participated to the Shot Boundary Detection (SBD) task and the Feature Extraction (FE) task (detection of person X only) of the TRECvid 2003 workshop.

2 Shot Boundary Detection Task

The system used by CLIPS-IMAG to perform the TRECvid SBD task is almost the same as the one used for the TREC-10 and TREC-11 video track evaluations [2] [1]. This system detects "cut" transitions by direct image comparison after motion compensation and "dissolve" transitions by comparing the norms of the first and second temporal derivatives of the images. It also has a special module for detecting photographic flashes and filtering them as erroneous "cuts" and a special module for detecting additional "cuts" via a motion peak detector. The precision versus recall or noise versus silence compromise is controlled by a global parameter that modify coherently the system internal thresholds. The system is still globally organized according to a (software) dataflow approach and Figure 1 shows its architecture.

The original version of this system was evaluated using the INA corpus and the standard protocol [3] (http://clips.imag.fr/mrim/-georges.quenot/OT10.3/aim1/) developed in the context of the GT10 working group on multimedia indexing of the ISIS French research group on images and signal processing. We partly reused this test protocol (with different test corpora) for the TREC-10, TREC-11 and TRECvid SBD tasks. The reference segmentation for the development and test collections of the TRECvid corpus were also built with this system (the version used for the TREC-11 evaluation).

Very little modification was made relatively to the TREC-11 version of the system, only minor adjustments of control parameter. The main additional work was an attempt to get a precise control of the noise to silence ratio.

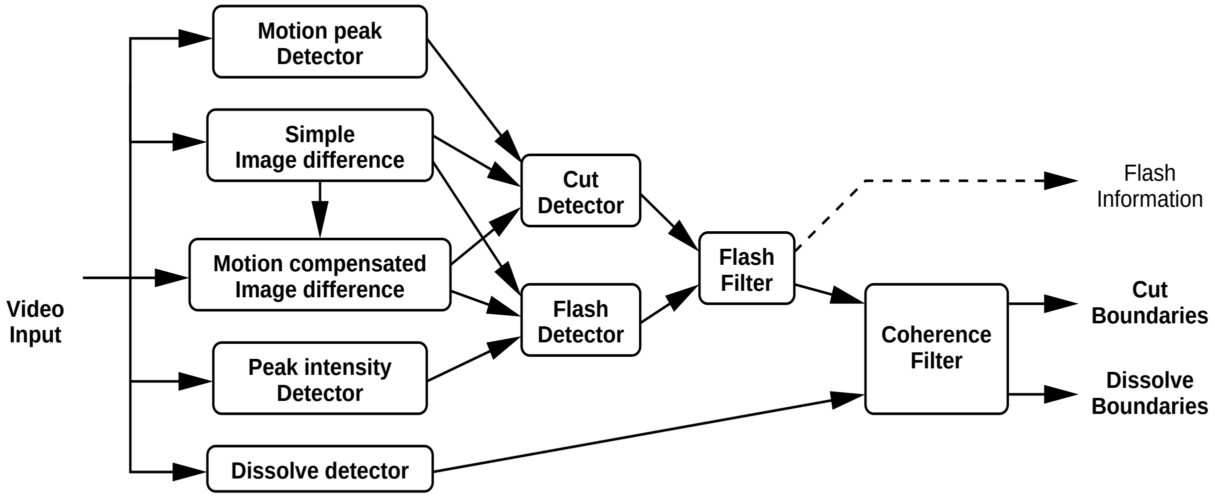

Figure 1: Shot boundary detection system architecture

2.1 Cut detection by Image Comparison after Motion Compensation

This system was originally designed in order to evaluate the interest of using image comparison with motion compensation for video segmentation. It has been complemented afterward with a photographic flash detector and a dissolve detector.

2.1.1 Image Difference with Motion Compensation

Direct image difference is the simplest way for comparing two images and then to detect discontinuities (cuts) in video documents. Such difference however is very sensitive to intensity variation and to motion. This is why an image difference after motion compensation (and also gain and offset compensation) has been used here.

Motion compensation is performed using an optical flow technique [4] which is able to align both images over an intermediate one. This particular technique has the advantage to provide a high quality, dense, global and continuous matching between the images. Once the images have been optimally aligned, a global difference with gain and offset compensation is computed.

Since the image alignment computation is rather costly, it is actually computed only if the simple image difference with gain and offset compensation alone has a high enough value (i.e. only if there is significant motion within the scene). Also, in order to reduce the computation cost, the differences (with and without motion compensation) are computed on reduced size images (typically 96×72 for the PAL video format). A possible cut is detected if both the direct and the motion compensated differences are above an adaptive threshold.

In order for the system to be able to find shot continuity despite photographic flashes, the direct and motion compensated image difference modules does not only compare consecutive frames but also, if needed, frames separated by one or two intermediate frames.

2.1.2 Photographic flash detection

A photographic flash detector feature was implemented in the system since flashes are very frequent in TV news (for which this system was originally designed for) and they induce many segmentation errors. Flash detection has also an interest apart from the segmentation problem since shots with high flash densities indicates a specific type of event which is an interesting semantic information.

The flash detection is based on an intensity peak detector which identify 1- or 2-frame long peaks of the average image intensity and a filter which uses this information as well as the output of the image difference computation modules. A 1- or 2-frame long flash is detected if there is a corresponding intensity peak and if the direct or motion compensated difference between the previous and following frames are below a given threshold. Flash information may be output toward another destination. In the segmentation system, it is used for filtering the detected “cut” transitions.

2.1.3 Motion peak detection

It was observed from TREC-10 and other evaluations that the motion compensated image difference was generally a good indicator of a “cut” transition but, sometimes, the motion compensation was too good at compensating image differences (and even more when associated to a gain and offset compensation) and quite a few actual “cuts” were removed because the pre- and post-transition images were accidentally too close after motion compensation. We found that it is possible not to remove most of them because such compensation usually requires compensation with a large and highly distorted motion which is not present in the previous and following image-to-image change. A “cut” detected from simple image difference is then removed if it is not confirmed by motion compensated image difference *unless* it also corresponds to a peak in motion intensity.

2.2 Dissolve detection

Dissolve effects are the only gradual transition effects detected by this system. The method is very simple: a dissolve effect is detected if the L_1 norm (Minkowski distance with exponent 1) of the first image derivative is high enough compared to the L_1 norm of the second image derivative (this checks that the pixel intensities roughly follows a linear but non constant function of the frame number). This actually detects only dissolve effects between constant or slowly moving shots. This first criterion is computed in the neighborhood (± 5 frames) of each frame and a filter is then applied (the effect must be detected or almost detected in several consecutive frames).

2.3 Output filtering

A final step enforces consistency between the output of the cut and dissolve detectors according to specific rules. For instance, if a cut is detected within a dissolve, depending upon the length of the dissolve and the location of the cut within it, it may be decided either to keep only one of them or to keep both but moving one extremity of the dissolve so that it occurs completely before or after the cut.

2.4 Global tuning parameters

The system has several thresholds that have to be tuned for an accurate detection. Depending upon their values, the result can detect or miss more transitions. These thresholds also have to be well balanced among themselves to produce a consistent result. Most of them were manually tuned as the system was built in order to produce the best possible results using development data.

For the TREC-11 and following evaluation,s as well as for other applications of the system, we decided to have all the threshold parameters be a function of a global parameter controlling the recall versus precision compromise (or, more precisely, the silence to noise ratio). We actually used two such global parameters: one for the cut transitions and one for the gradual transitions. A function was heuristically devised for each system threshold for how it should depend upon the global parameters.

Ten values were selected for the global parameters. These values were selected so that they cover all the useful range (outside of this range, increasing or decreasing further the global parameter produces a loss on both the silence and noise measures) and within that range they set targets on a logarithmic scale for the silence to noise ratio. For the cut transitions, the target for the base 2 logarithm of silence/noise ranged from -5.0 to +4.0 with a step of 1.0. or the gradual transitions, the target for the base 2 logarithm of silence/noise ranged from -1.0 to +1.25 with a step of 0.25. The values for the target ratios as well as the target range were obtained by tuning the system global control parameters on the TREC 2001 Shot Boundary Detection test collection.

2.5 Results

Ten runs have been submitted for the CLIPS-IMAG system. These correspond to the same system with a variation of the global parameter controlling the silence versus noise compromise.

As expected, this made possible the drawing of a recall × precision curve. Figure 2 shows these curves for the features selected for the evaluation. There are three recall × precision curves respectively for all transitions, for cut transitions and for gradual transitions. There is also a frame-recall × frame-precision curve that qualifies the accuracy of the boundaries of recovered gradual transitions. For comparison purposes, the results of other systems are plotted as set of points (with abbreviated names given with the results by NIST).

The silence to noise ratio targets were missed by a large amount indicating that the TRECvid 2003 SBD collection is very different in content from the TREC 2001 one. The overall performance of all SBD systems is also much less on the TRECvid 2003 SBD collection. This is probably due to a large amount of special effects and highly dynamic visual jingles that induces a lot of false positive. Also, the TV news include content with more motion that the relatively static TREC 2001 corpus.

The CLIPS system appears to be quite good for gradual transitions both for their detection and location. This indicates that the chosen method (comparison of the first and second temporal derivative of the images) is quite good even if theoretically suited only for sequences with no or very little motion. The CLIPS system is third for cut detection and second for gradual transition detection, for gradual transition location, and for all transitions.

3 Feature Search Task

CLIPS extracted only features 27 “Person X = Madeleine Albright”. We did the detection only from the processing of the audio track.

3.1 Person X Detection by speaker identification

3.1.1 Parameterization of the audio signal

We used 16 MFCC (Mel Frequency Cepstral Coefficients) coefficients and the log energy computed every 10 ms on 20 ms windows with no Cepstral Mean Subtraction (CMS) applied. We used GMMs (Gaussian Mixtures Models) to characterize the speaker X models. The GMMs where made of 128 gaussian distributions with diagonal covariance matrix per distribution and were trained using the ELISA platform [5].

3.1.2 Decision

The idea was to train a Person X speech model namely a Madeleine Albright model and a world speech model (corresponding to nonX model) and to compute the log-likelihood ratio between both models. For the person X model we used all we could manually find in the dev corpus (about 90 seconds) and we also used external data from web and Hub4 96 corpus. For the world model we used the entire dev corpus except segments containing the person X.

For person X model since there was not enough data in the dev corpus we trained the model by adaptation of an existing model. Basically we used a female model trained on about 2 hours of clean studio speech that we adapt by MAP (Maximum A Posteriori) adaptation on the available person X data.

Suppose we have a person X model M_X, a world model M_{nonX} and an unknown acoustic vector sequence $S = s_1 \ldots s_n$. The log-likelihood ratio between the hypothesis of S being uttered by the person X and not, is defined by:

$$llr(S) = \log P(S/M_X) - \log P(S/M_{nonX})$$

The bigger the ratio is the bigger the probability that the sequence S was uttered by X. The log-likelihood ratio was computed for every entire shot

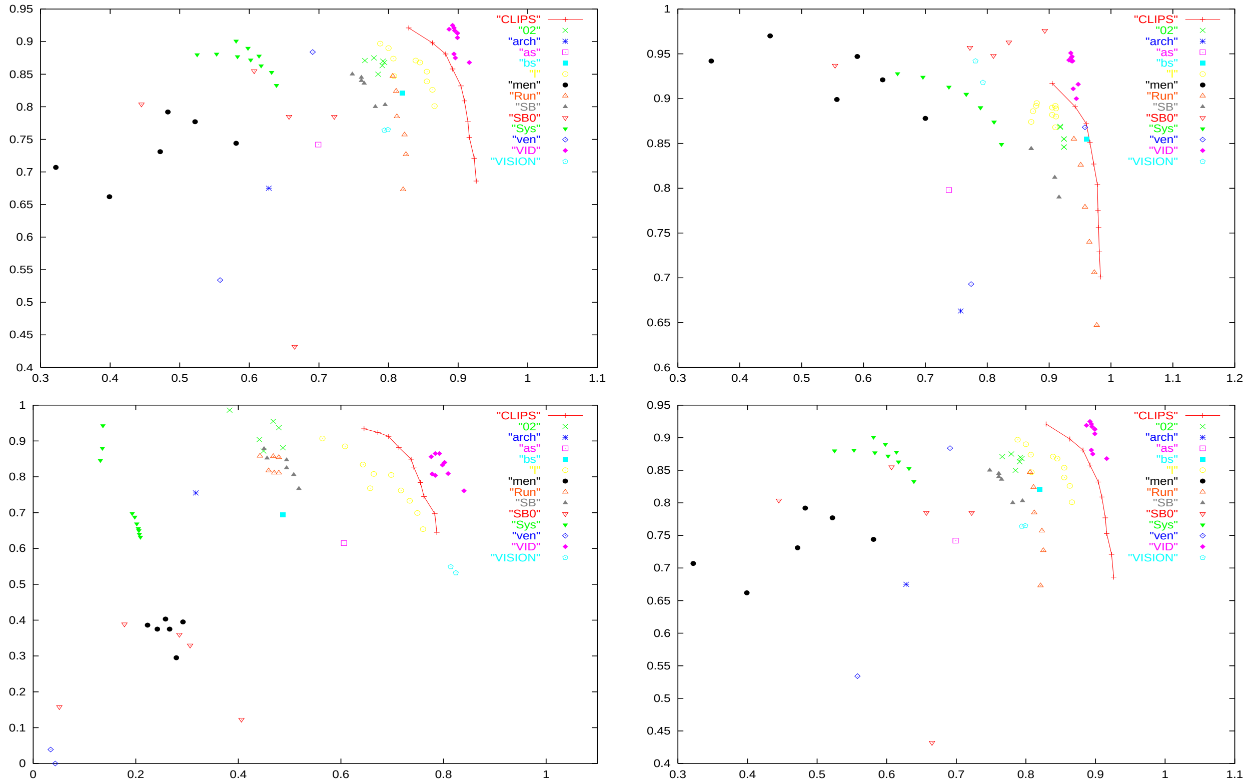

Figure 2: Recall × Precision global results for all (top left), cut (top right) and gradual (bot. left) transitions; Frame-Recall × Frame-Precision global results for gradual transitions (bot. right).

and the results were sorted descendant. Since a shot must contain speech in order to be selected we used the LIMSI speech transcriptions in order to eliminate the shots that did not contain enough speech. Thus, from an initial total of 35220 shots we kept 30587 shots containing at least 60 % of speech signal in our experiments.

3.1.3 Results

The test collection was containing 42 person X (Madeleine Albright) shots. Unfortunately among those shots she was only speaking in 5 which explains the poor average precision results we finally obtained on person X feature. Moreover, those five shots, where M. Albright speaks, obtained respectively the following positions: 4, 190, 365, 1380, 1704 in the ordered list of 30587 "candidate" shots. Again, these results look like disappointing and may be due to mismatch between the raining data (to learn person X model) and the test audio conditions. For instance, the 2 shots that had the worst results were respectively an exterior shot and a shot were more than one person is speaking.

It is also important to note that in a traditional speaker verification task (for instance in NIST speaker verification evaluation protocols [6] [7]), the ratio between client (person X) speakers and impostors (non X) is generally 1 per 10 or 1 per 20. Here, in the TREC framework, the ratio becomes 1 per 6000 (5/30587) which changes completely the difficulty of the task ! Looking again at our results, we can note that the 5 person X shots were all ranked in the first 6 % (1704/30587) of the total shots.

3.2 Person X Detection from audio transcription

We also try to perform the Person X Detection task by using the audio transcription. We simply selected shots for which the name of Person X appeared in the transcription provided by LIMSI using their broadcast news transcription system [9]. We selected these shots plus the nine following shots with a progressively decreasing priority.

We found 19 out of the 42 relevant shots at a recall of 100 and 23 of them at a recall of 1000 for a global average precision of 0.129.

We also tried several combinations of detection form person X voice and detection from person X name in the transcription but since detection of person X voice performed very poorly, all combinations were less good that detection from person X name in the transcription alone.

4 Conclusion

We have presented the systems used by CLIPS-IMAG to perform the Shot Boundary Detection (SBD) task and the Feature Extraction (FE) task of the TRECvid workshop. Results obtained for the 2003 evaluation were presented. The CLIPS SBD system based on image difference with motion compensation and direct dissolve detection was second among 14 systems. This system gives control of the silence to noise ratio over a wide range of values and for an equal value of noise and silence (or recall and precision), the value is 12 % for all types of transitions. Detection of person X from speaker recognition alone was deceiving due to the small number of shots containing person X in the overall test collection (about 1/700) and the even small number in which person X was actually speaking (about 1/6000). Detection of person X from speech transcription performed much better but was still lower than other systems using also the image track for the detection.

References

[1] Quénot, G.M.: CLIPS at TREC-11: Experiments in Video Retrieval, In em 11th Text Retrieval Conference, Gaithersburg, MD, USA, 19-22 November, 2002.

[2] Quénot, G.M.: TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation, In em 10th Text Retrieval Conference, Gaithersburg, MD, USA, 13-16 November, 2001.

[3] Ruiloba, R., Joly, P., Marchand, S., Quénot, G.M.: Toward a Standard Protocol for the Evaluation of Temporal Video Segmentation Algorithms, In *Content Based Multimedia Indexing*, Toulouse, Oct. 1999.

[4] Quénot, G.M.: Computation of Optical Flow Using Dynamic Programming, In *IAPR Workshop on Machine Vision Applications*, pages 249-52, Tokyo, Japan, 12-14 nov. 1996.

[5] Magrin-Chagnolleau, I., Gravier, G, Blouet, R. for the ELISA consortium: Overview of the 2000-2001 ELISA consortium research activities, In *2001: A Speaker Odyssey*, pp.6772, Chania, Crete, June 2001.

[6] Przybocki, M., Martin, A.: NIST's Assessment of Text Independent Speaker Recognition Performance, The Advent of Biometrics on the Internet, A COST 275 Workshop in Rome, Italy, Nov. 7-8 2002

[7] Moraru, D., Meignier, S.,Besacier, L., Bonastre, J.-F., Magrin-Chagnolleau, Y.: The ELISA Consortium Approaches in Speaker Segmentation during The NIST 2002 Speaker Recognition Evaluation In *Proceedings of ICASSP*, Hong Kong, 6-10 apr. 2003.

[8] Delacourt, P., Wellekens, C.: DISTBIC: a speaker-based segmentation for audio data indexing, In *Speech Communication*, Vol. 32, No. 1-2, September 2000.

[9] Gauvain, J.L., Lamel, L., Adda, G.: The LIMSI Broadcast News Transcription System, In *Speech Communication,* 37(1-2):89-108, May 2002.