TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation

Georges M. Quénot

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

Abstract

This paper presents the system used by CLIPS-IMAG to perform the Shot Boundary Detection (SBD) task of the Video track of the TREC-10 conference. Cut detection is performed by computing image difference after motion compensation. Dissolve detection is performed by the comparison of the norm over the whole image of the first and second temporal derivatives. An output filter is added in order to clean the data of both detector and to merge them into a consistent output. This system also has a special module for detecting photographic flashes and filtering them as erroneous "cuts". Results obtained for the TREC-10 evaluation are presented. The system appear to perform in a very good way for cut transitions and within the average for gradual transitions.

1 Introduction

Temporal video segmentation has very important applications in video document indexing and retrieval, in information or emission type filtering and in video document browsing among many others. It must be distinguished from spatial (extracting objects) and spatio-temporal (tracking objects) segmentations that will not be considered here (even though probably also useful for video document indexing). This work focuses solely on the segmentation of the image track of video documents. The segmentation process consists mainly in detecting "transition effects" between "homogeneous segments" (shots), the definition of which is rather application-dependent. Transition effects may be roughly classified into three categories:

- "cuts": sharp transitions between two consecutive images, the second image is completely or almost completely different from the first one,
- "dissolves": continuous transition between two continuous sequences by a progressive linear combination of them (this includes "fades in" and "fades out"),
- "others": all other type of transitions, including all possible special effects.

Several levels of difficulty arise within the global segmentation task:

- The most easy and low level is to find "cuts" and "dissolves" between almost static images. Specific filters can be quite easily designed for this task.
- More difficult is the detection of "dissolves" in the general case and other effects (such as wipes for instance). Higher level tools can also be developed for detecting such difficult to find transitions or as well these and the simple ones simultaneously [1].
- Finally, the highest level of difficulty is to find among the identified transitions or segments, which of them are significant at the semantic level (possibly hierarchically) in order to be able to structure the document [2].

The transition effect definition is not always straightforward and may depend upon the target applications. For instance, it has been decided in our case that cuts, even obvious, appearing inside "visual jingles" and stroboscopic effects should not be counted as actual cuts. All effects are counted only if they correspond to a transition for the whole image. Superimposed text, small images and logos appearance and disappearance are not counted as transition effects.

Many automated tools for the temporal segmentation of video streams have been already proposed. It is possible to find some papers that are providing state of the art of such methods [3] [4] [5]. In this paper, we describe the temporal video segmentation system used by CLIPS-IMAG to perform the Shot Boundary Detection (SBD) task of the Video track of the TREC-10 conference. This system was first developed at the LIMSI-CNRS laboratory and was then improved at the CLIPS-IMAG laboratory. It 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. This system also has a special module for detecting photographic flashes and filtering them as erroneous "cuts". It is globally organized according to a (software) dataflow approach and Figure 1 shows its architecture.

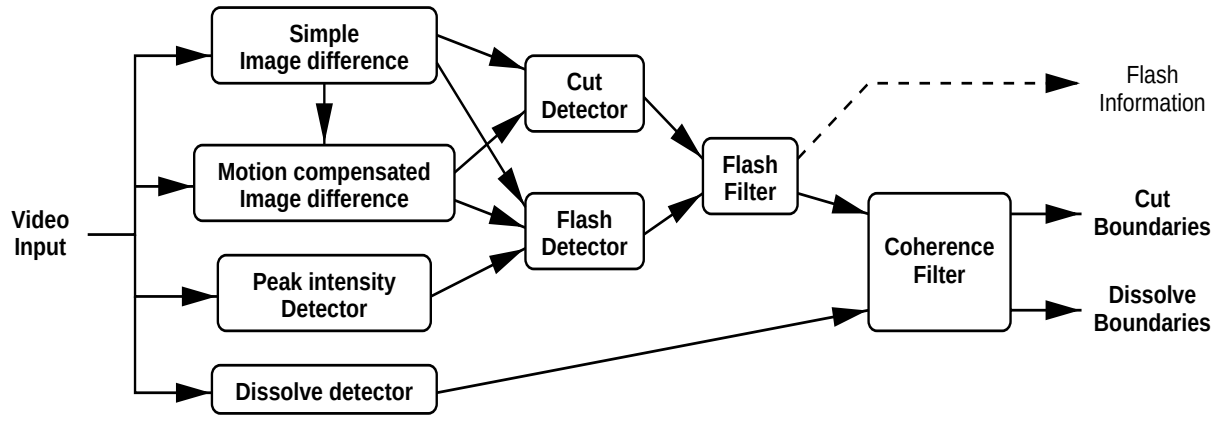

Figure 1: System architecture

The original version of this system was evaluated using the INA corpus and the standard protocol [6] (`http://asim.lip6.fr/AIM/corpus/aim1/indexE.html`) developed in the context of the GT10 working group on multimedia indexing of the ISIS French research group on images and

signal processing. The TREC-10 SBD task partly reused this test protocol (with a different test corpus).

2 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 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 [7] 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.

2.2 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 density 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.

3 Dissolve detection

Dissolve effects are the only continuous 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 ($\pm$ 3 frames) of each frame and a filter is then applied (the effect must be detected or almost detected in several consecutive frames).

4 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.

The system is designed for having the capability to evolve by including within its dataflow architecture new feature detection modules and new decision modules. It may also output other data than segmentation information like detection of photographic flashes or other features.

5 Evaluation using the TREC test data

The results for two variants of the CLIPS system were submitted for the TREC SBD task. These variants differ only in the value of some control (threshold) parameters. They are laleled "CLIPS-1" and "CLIPS-2" (CL-1 and CL-2 in the tables) respectively. The first one corresponds to the original parameters of the system (which was tuned for French TV news segmentation). The second one was set with lower thresholds in order to try a configuration with a higher recall, possibly resulting also into a lower precision. The threshold parameters were changed only for the cut detection part. None of the threshold parameters were tuned using the part of the TREC-10 corpus, neither with the data used for system evaluation, nor with the data unused for system evaluation. Shot boundary detection was performed on all of the test data specified for the TREC-10 SBD task in from 5 to 10 times real time (using a Pentium III @ 800 MHz), depending upon the documents' content.

The results are presented on the basis of the final version of reference data and comparison software which give slightly different results from the draft version. Only the system which provided results for the whole test set are compared with our system. These include two systems from Fudan University, China (FU-1 and FU-2), two systems from IBM Almaden Research Center, USA (IBM-1 and IBM-2), one system from Imperial College - London, UK (ICKM), one system from Microsoft Research, China (MSSD), one system from Glasgow University, UK (MB_Frequency, MBF), and two systems from University of Amsterdam and TNO, the Netherlands (Media Mill, MM-1 and MM-2). Global deletion and insertion rates, recall and precision on all files (2061 cuts, 1108 gradual,

3169 total transitions for a total of 624267 frames in 42 video documents) are used as a synthetic data and are presented in table 1.

Cuts	**CL-1**	**CL-2**	FU-1	FU-2	IBM-1	IBM-2	ICKM	MBF	MSSD	MM-1	MM-2
Del.	**0.012**	**0.011**	0.030	0.030	0.021	0.035	0.065	0.184	0.072	0.053	0.091
Ins.	**0.105**	**0.293**	0.040	0.040	0.038	0.037	0.112	0.568	0.074	0.358	0.117
Rec.	**0.988**	**0.989**	0.970	0.970	0.979	0.965	0.935	0.816	0.928	0.947	0.909
Pre.	**0.904**	**0.771**	0.961	0.961	0.963	0.963	0.893	0.590	0.926	0.726	0.886
Grad.	**CL-1**	**CL-2**	FU-1	FU-2	IBM-1	IBM-2	ICKM	MBF	MSSD	MM-1	MM-2
Del.	**0.293**	**0.291**	0.379	0.402	0.268	0.230	0.360	0.963	0.306	0.222	0.554
Ins.	**0.566**	**0.565**	0.241	0.214	0.447	0.589	0.433	0.000	0.375	0.388	0.067
Rec.	**0.707**	**0.709**	0.621	0.597	0.732	0.770	0.640	0.037	0.694	0.778	0.446
Pre.	**0.555**	**0.555**	0.72O	0.736	0.621	0.566	0.596	1.000	0.649	0.667	0.870
All	**CL-1**	**CL-2**	FU-1	FU-2	IBM-1	IBM-2	ICKM	MBF	MSSD	MM-1	MM-2
Del.	**0.110**	**0.109**	0.152	0.160	0.107	0.103	0.168	0.457	0.154	0.112	0.253
Ins.	**0.266**	**0.388**	0.110	0.101	0.181	0.230	0.224	0.369	0.179	0.368	0.100
Rec.	**0.890**	**0.891**	0.848	0.840	0.893	0.897	0.832	0.543	0.846	0.888	0.747
Pre.	**0.770**	**0.697**	0.885	0.893	0.832	0.796	0.788	0.595	0.825	0.707	0.882

Table 1: Global results for the SBD TREC-10 evaluation, deletion and insertion rates, precision and recall for "cut", gradual and all transitions.

Table 1 results shows that our attempt to increase the recall (or decrease the deletion rate) of the "cut" transitions by reducing the thresholds between the variants CL-1 and CL-2 of our system completely failed while the precision was severely decreased (or the insertion rate severely increased). Also, the ratio between insertions and deletions is of 9:1 for CL-1 and of 27:1 for CL-2, which is highly asymmetrical. The reason is probably that our system CL-1 was already a highly "recall oriented" system designed to minimize the deletion rate while keeping the insertion rate reasonable (this choice was justified by the hypothesis that over-segmentation can be identified and removed in further steps and may not be very penalizing in most applications while, once missed, transitions cannot easily be detected again and their miss may be penalizing for applications). However such a ratio was not expected (CL-1 was tuned for about a 5:1 ratio on French TV news) and neither was the absence of any improvement in the insertion rate (or recall). The results show that the transitions missed by our system cannot be recovered with the approach used whatever the threshold choice. However, for both variants, the deletion rate is very low (about 1 %). It is about twice lower than the one of the following best system (IBM-1 with about 2 %) and four times lower that the average of all systems.

For gradual transitions, our system shows roughly a 3:1 insertions to deletions ratio (table 1). However, our system is designed to detect only "dissolve" gradual transitions and the deletion rate relative to "dissolve" transitions alone might be lower and, therefore, the actual ratio higher. There is little difference between CL-1 and CL-2 systems since there is no change in the thresholds for the dissolve detector. The minor difference comes from indirect effects of differences in cut detection via the output filter. The performance of CL-2 appear to be slightly better for CL-2 but the overall performance is much better for CL-1.

The insertion and deletion rates for all systems appear to be much higher for gradual transitions

than for cuts for all systems (a 10:1 ratio typically). Since cuts are only about twice as numerous as gradual transitions, the effect of errors on gradual transitions strongly dominates in the errors for all transitions (table 1).

Systems are hard to compare since the results include two independent measures: insertion and deletion rates (or precision and recall) and systems have an extremely variable insertions to deletions ratio: from 9:1 for CL-1 (excluding CL-2) for CL-1 down to 1:1 for MSSD (for cuts). However, deletion rates should be compared for an equivalent insertion rate or vice versa or, alternatively, they should be compared at a point for which both values are identical or in a pre-defined ratio. None of these performance indexes is significant without considering simultaneously the other independent variable. Currently, the TREC-10 SBD result data does not provide any single synthetic measure allowing to rank the systems.

Ruiloba et al. [6] proposed three different global indexes: the "error rate" which is the sum of insertion and deletion rates, "quality" which is equivalent to a weighted sum of them giving more importance to deletions than to insertions and "correction probability" which has the drawback of giving a lot more importance to deletions than to insertions, weighting them respectively with the total number of frames minus the number of transitions and the number of transitions alone. All of these measures have their bias and none was selected for TREC-10 SBD evaluation. Moreover, the one chosen would have to be known in advance so that the systems can be tuned appropriately (in terms of precision versus recall compromise) to it.

The best solution would have been that results be given for all systems for a wide range of insertions to deletions ratios (by varying internal threshold parameters) in order to produce a sound Recall $\times$ Precision curve. This would have allowed a more objective system performance comparison using for instance: precision at a given recall, recall at a given precision, or any of them for a fixed precision to recall ratio (or similar indexes using insertion and deletion rates instead of precision and recall). This was not possible because the test framework did not permit to provide a ranked list of detected transitions and allowed only two system output per institution.

We did, however, run our system with many different parameter sets (by varying only one global parameter, according to which all other vary simultaneously), we evaluate each run with the same software and reference data and were able to draw Recall $\times$ Precision and Deletion rate $\times$ Insertion rate diagrams which permitted to compare our system to all others. Both the cuts and global transition control parameters were varied here unlike in the two officially submitted runs into which ony the cuts control parameters were varied. Figures 2 and 3 shows on the same diagram the curve obtained by varying the CLIPS-IMAG system parameters and the points corresponding to all other systems. Figures 4 and 5 shows more detail results in the Deletion $\times$ Insertion plane.

From these data, it appears that when comparing the CLIPS system to the nine other systems (or the six other systems if we take only the best one for each institution that submitted two runs), it ranks:

- 2nd of ten (respectively 2nd of seven) for cuts,
- 5th of nine (respectively 3rd of six) for gradual transitions,
- 3rd of ten (respectively 2nd of seven) for all transitions,

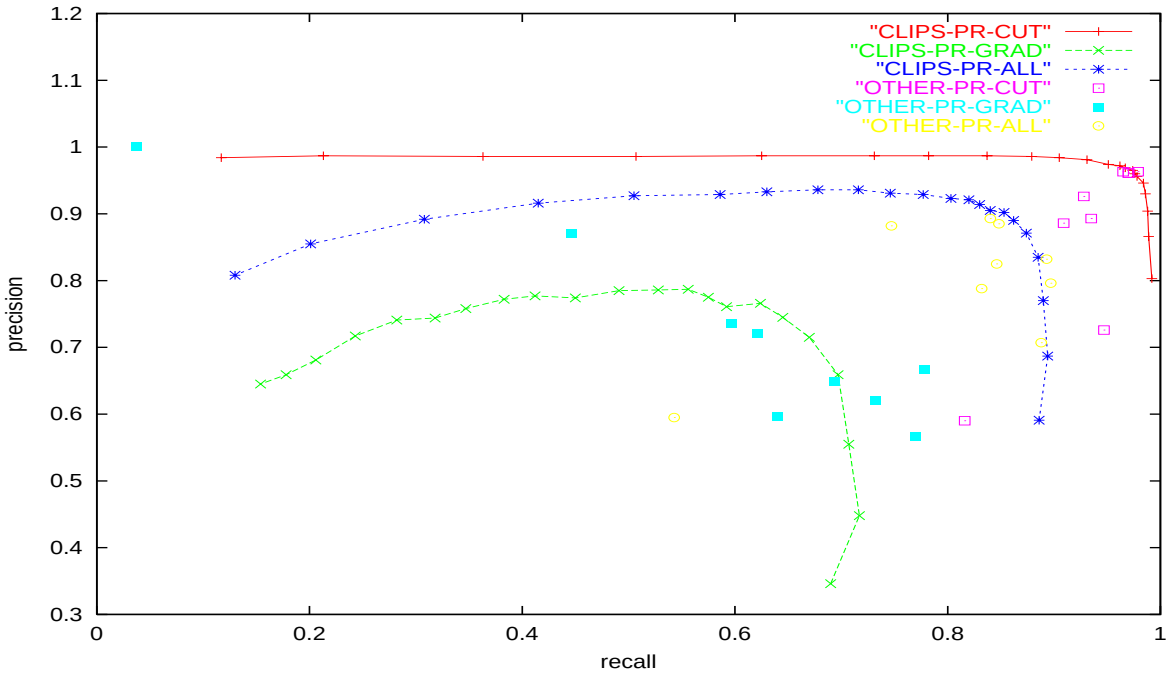

Figure 2: Global results: Recall × Precision

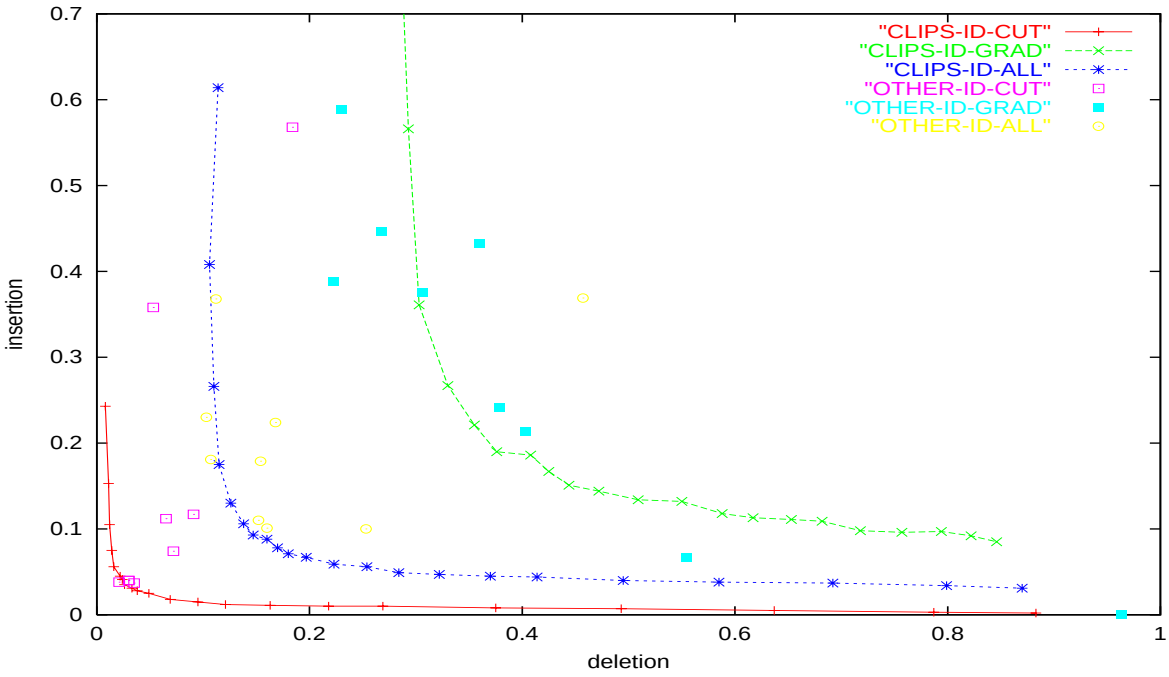

Figure 3: Global results: Deletion × Insertion

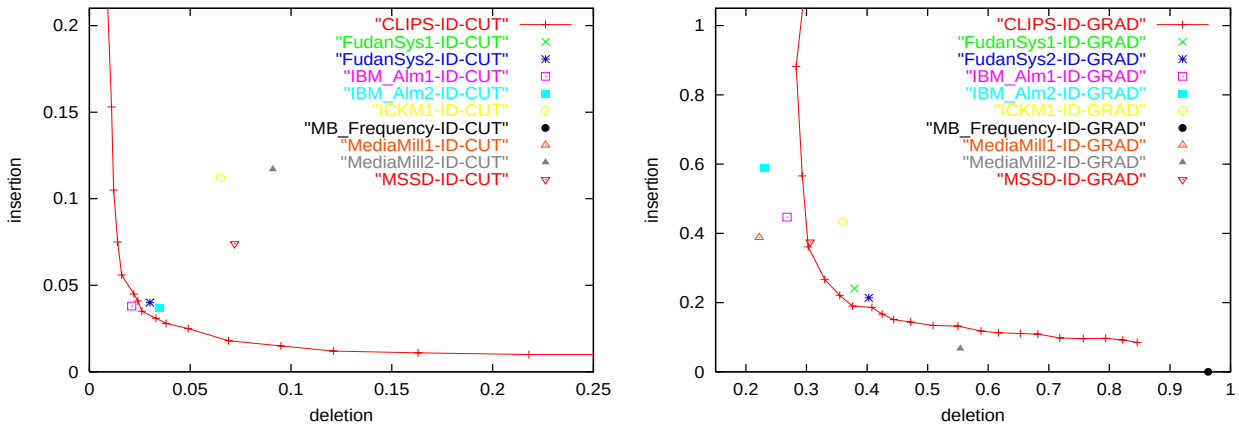

Figure 4: Global results for cuts (left) and gradual transitions (right) : Deletion × Insertion

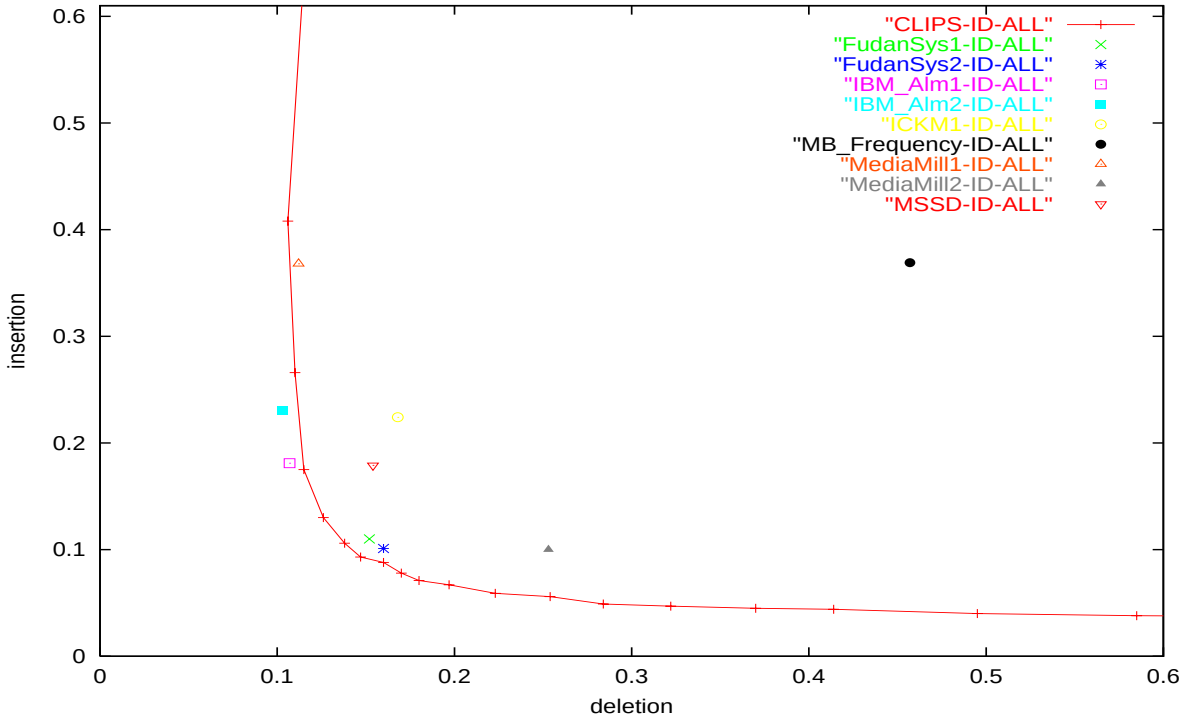

Figure 5: Global results for all transitions: Deletion × Insertion

The Recall × Precision and Deletion × Insertion curves do not appear to be monotonous as they usually are. This is because they are not obtained from a ranked list of detected transition but rather by modifying a set of parameters according to a single global control one. The combined effect of these various parameters, each controlling subsystems that interact with each other for adding or removing transition, explains these unusual results, possibly indicating a non optimal dependence of the several parameters from the global control one. For extreme values of the global

control parameter, there is a loss both on recall and precision simultaneously, possibly indicating also unrealistic conditions of operation for the system.

6 Conclusion

This paper has presented the system used by CLIPS-IMAG to perform the Shot Boundary Detection (SBD) task of the Video track of the TREC-10 conference. It implements cut detection using image difference after motion compensation and dissolve detection by the comparison of the norm over the whole image of the first and second temporal derivatives. It also incorporate an output filter to clean the data of both detector and to merge them into a consistent output, and a special module for detecting photographic flashes and filtering them as erroneous "cuts". Shot boundary detection was performed on all of the test data specified for the task in from 5 to 10 times real time, depending upon the documents content. The CLIPS system appear to perform in a very good way for cut transitions and in the average for gradual transitions.

References

[1] Lin,Y., Kankanhalli, M., Chua, T.-S.,: Temporal multi-resolution analysis for video segmentation, In *ACM Multimedia Conference*, 1999.

[2] Aigrain, P., Joly, P., Longueville, V.,: Medium-Knowledge-Based Macro-Segmentation of Video into Sequences, In *Intelligent Multimedia Information Retrieval, Ed. Mark MAYBURY, AAAI Press, MIT Press*, pp 159–173, 1997

[3] Aigrain, P., Zhang, H.J., Petkovic, D.,: Content-based representation and retrieval of visual media : a state-of-the-art review, In *Multimedia Tools and Applications*, 3(3):179-202, November 1996

[4] Lupatini, G., Saraceno, C., Leonardi, R.: Scene break detection: a comparison, In *Proc. VIIIth Intern. Workshop on Research Issues in Data Engineering: Continuous-Media Databases and Applications*, 34-41, Feb. 98.

[5] Boreczky, J. S., Rowe, L. A.: Comparison of video shot boundary detection technique. In *IS&T/SPIE Conference on Electronic Imaging Technology and Science*, San Jose, USA, February 1996.

[6] 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.

[7] 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.