CLIPS at TREC-11: Experiments in Video Retrieval

Georges M. Quénot, Daniel Moraru, Laurent Besacier and Philippe Mulhem

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, the Feature Extraction (FE) and the Search (S) task of the Video track of the TREC-11 conference. Results obtained for the TREC-11 evaluation are presented.

1 Introduction

The CLIPS-IMAG laboratory has participated to all of the three tasks proposed in the video track of the TREC-11 evaluation. This participation was done in collaboration with teams from other institutions including LIMSI-CNRS (Orsay, France) for speech transcription, LIT-IPAL (Singapore) for face detection and INSA (Lyon, France) for text transcription. The following sections describe our participation to the tasks.

2 Shot Boundary Detection Task

The system used by CLIPS-IMAG to perform the TREC-11 SBD task is almost the same as the one used for the TREC-10 evaluation [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". With respect to the system used for the TREC-10 evaluation, this one has an additional module for detecting additional "cuts" via a motion peak detector. Some parameters controlling the existing modules have been tuned using the TREC-10 SBD corpus and reference segmentation, and a global parameter for the tuning of the recall versus precision compromise has been inserted. 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 [2] (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 and TREC-11 SBD tasks partly reused this test protocol (with different test corpora). The reference segmentation for the search, the feature test and the feature search collections of the TREC-11 corpus were also built with this system (the version used for the TREC-10 evaluation).

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 [3] 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.

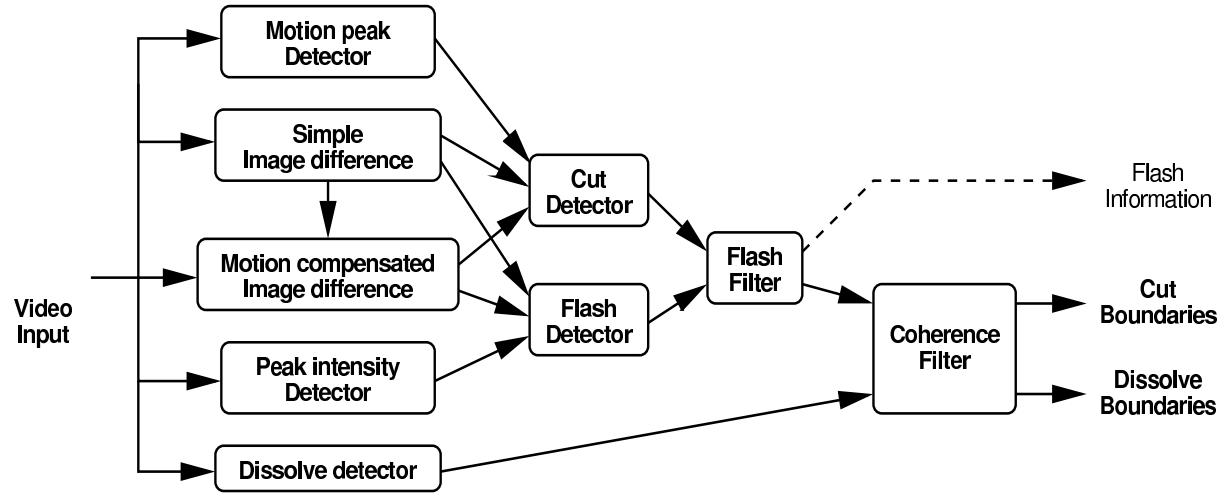

Figure 1: Shot boundary detection system architecture

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 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.

2.2 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 ($\pm$ 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 New features

2.4.1 Motion peak detection

The main new feature of the system is the motion peak detection module. 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 wich 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.

2.4.2 Global tuning parameter

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 sample data. No additional tuning was done for the TREC-10 evaluation. A first run was made using the default system threshold (originally oriented toward a high recall) and a second run with lower thresholds (20 % lower) in order to further improve the recall.

For the TREC-11 evaluation, 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 false positive to false negative ratio). A function was heuristically devised for all of them. A power low has been chosen. A first system tuning was done using the TREC-10 SBD corpus and reference segmentation in order to set a point at which the false positives are roughly equivalent to the false negatives. Then a power coefficient has also been tuned for each parameter in order to have the ratio to follow also roughly a power law.

2.5 Evaluation using the TREC-11 SBD test data

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 recall versus precision compromise. This parameter has been varied so that the target false positive to false negative ratio has extreme values of roughly 3:1 and 1:3 with intermediate ones following roughly a power law.

As expected, this made possible the drawing of a recall $\times$ precision curve. Figure 2 shows these curves for the features selected for the evaluation. There are three recall $\times$ precision curves respectively for all transitions, for cut transitions and for gradual transitions. There is also a frame-recall $\times$ 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 CLIPS system appears to be very good for gradual transitions both for the detection and the location. This may come from the specificity of TREC-11 video data which are quite old and which mostly contain dissolve or fade gradual transitions (other special effects were not common in the forties/fifties). This is the only type of gradual effect our system was designed for. This indicates also 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 appears to be in the average for cut detection but thanks to its very good performance in gradual transition detection and considering that these are more difficult to detect than cuts, its global performance for all transitions also remains very good.

3 Feature Search Task

CLIPS extracted only features 3 (faces), 4 (people), 8 (speech) and 10 (monologue).

3.1 Face and People Detection

Face and people detection were based on a face detection tool available from CMU [4]. This tool was run (by Philippe Mulhem and colleagues at Laboratories for Information Technologies, Singapore) on one keyframe automatically extracted for each shot. The keyframe was selected within the shot simply as the one having the highest contrast (in order to avoid frames within

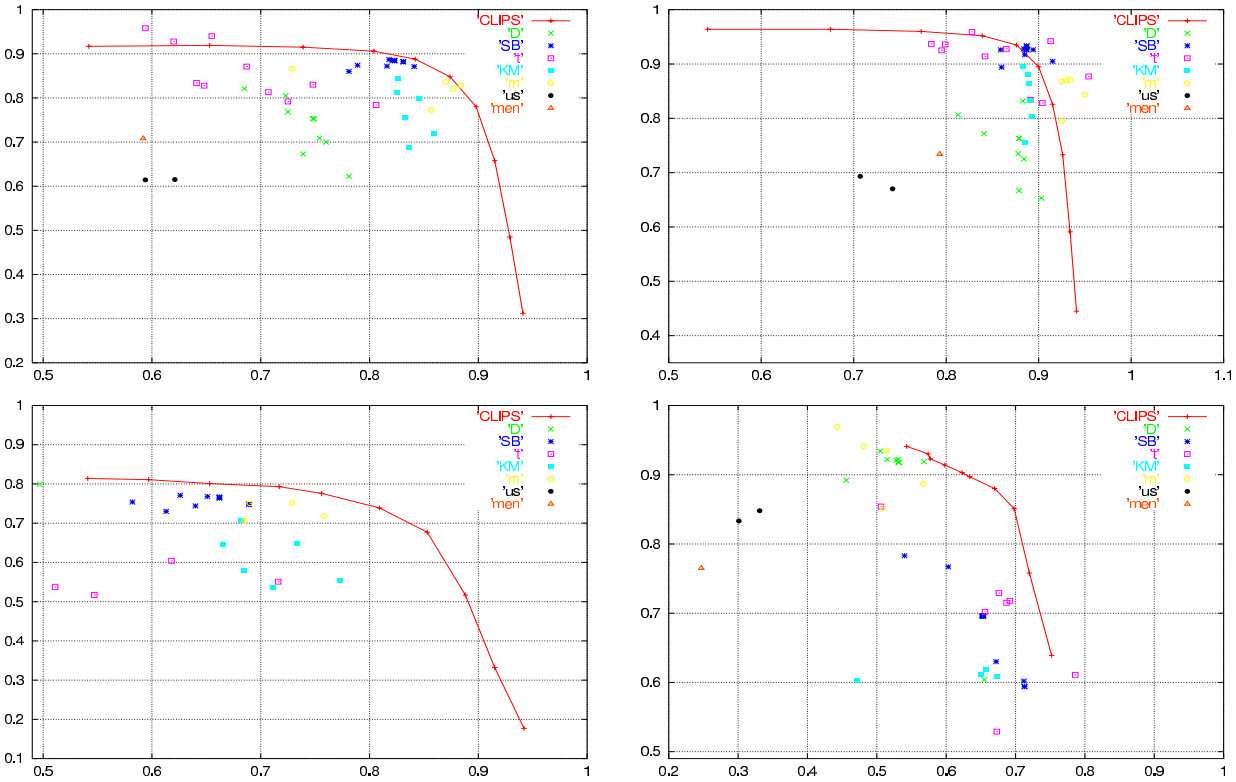

Figure 2: Recall × Precision global results for all (top left), cut (top right) and gradual (bot. left) transitions; Frame-Recall × Frame-Precision global results for gradual transitions (bot. right).

fades and dissolves). People were only detected on the basis of the presence of at least two faces. The results were ranked according to the presence of one (or at least two) face(s) and to the total face area.

Table 1 and 2 show the performance of the CLIPS system among other systems that have searched for features 3 and 4. The quality is quite low for these features. This comes probably from the simplicity of the approach only based on keyframe extraction followed by face detection (which is by itself quite good however), especially for people detection.

3.2 Speech and Monologue Detection

3.2.1 Speech Feature Detection

Both for speech and monologue feature detection, the acoustic vectors extracted from speech were conventional parameters used in speech processing, i.e. 16 MFCC (Mel Frequency Cepstral Coefficients) and their log energy computed every 10ms on 20 ms signal windows with no Cepstral Mean Subtraction (CMS) applied.

At first, we eliminated the silent films using the energy calculation of the signal. Then, the idea (Figure 3) was to train a speech model and a non-speech model (also called world model) and to compute the log-likelihood ratio between both models. We used GMMs (Gaussian Mixtures Models) to characterize speech and non-speech. The GMMs were made of 128 gaussian components and trained using the ELISA platfrom [5].

Suppose we have speech model M_{Spch}, a world model M_{UnSpch} and a acoustic vector sequence $X = x_1 \dots x_n$. The log-likelihood ratio between the hypothesis of X being speech and not being speech is defined by:

$$llr(X) = \log P(X/M_{Spch}) - \log P(X/M_{UnSpch})$$

The bigger the ratio is the bigger the probability of X being speech is.

The speech model M_{Spch} was trained on about 2.5 hours of speech manually selected from the DEV files. The world model M_{UnSpch} was trained on everything that was left from the DEV files (about 2.5 hours). The log-likelihood ratio was computed for every shot and the results were sorted descendant.

rank	system	A.P.	D.100	D.1000	rank	system	A.P.	D.100	D.1000
1	B_r1_1	0.613	99	303	6	B_E2002_1	0.154	53	114
2	B_RA_1	0.473	86	253	7	B_om1_1	0.150	28	255
3	B_M-1_1	0.327	51	312	8	B_Sys1_1	0.111	17	190
4	B_M-2_2	0.288	53	293	9	B_l2_2	0.091	56	57
5	**CLIPS**	**0.178**	**70**	**118**	10	B_l1_1	0.089	55	55

Table 1: Average precision and average hits at depth 100 and 1000 for feature 3 (faces).

rank	system	A.P.	D.100	D.1000	rank	system	A.P.	D.100	D.1000
1	A_r2_2	0.274	57	277	6	B_r1_1	0.050	45	48
2	B_M-1_1	0.271	31	361	**7**	**CLIPS**	**0.023**	**18**	**18**
3	B_T1_1	0.248	54	251	8	B_l1_1	0.008	12	12
4	B_T2_2	0.168	27	223	9	B_l2_2	0.008	10	10
5	B_Sys1_1	0.071	44	83					

Table 2: Average precision and average hits at depth 100 and 1000 for feature 4 (people).

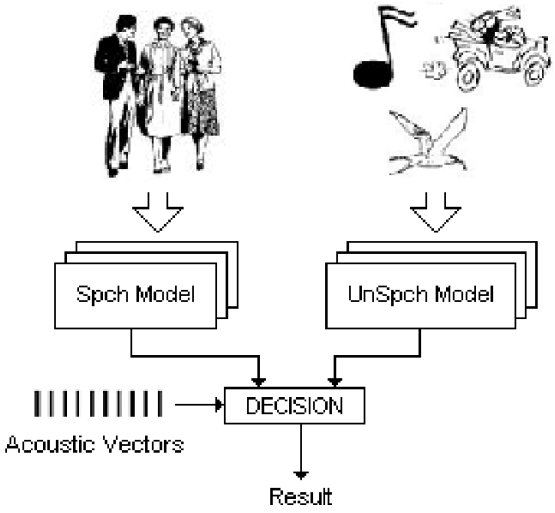

Figure 3: Speech Feature Detection

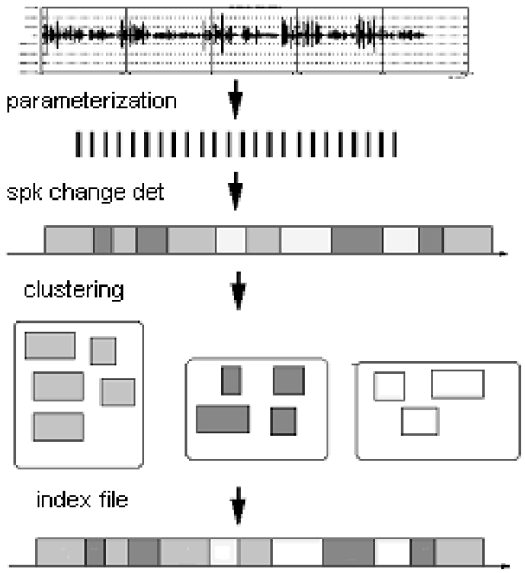

Figure 4: Speaker Segmentation System

3.2.2 Monologue Feature Detection

For the monologue feature detection we used the CLIPS Segmentation System [6] used during last NIST 2002 Speaker Recognition Evaluation, combined with the results from the speech feature detection task. The CLIPS Speaker Segmentation System is presented in Figure 4.

Once the speech is parameterized the segmentation is done in two steps. At first the speaker change points are detected using the Bayesian Information Criterion [7]. The purpose is to cut the file in single-speaker segments. Then the segments are grouped by speakers using an hierarchical clustering algorithm. At the end an index file is created containing all the speaker information obtained.

In order to perform monologue detection only on speech segments, the speaker segmentation system was applied only on the TEST files that had at least one shot in the top 300 of the speech feature detection task. Then, the shots labeled as monologue shots where the shots which were found to contain only one speaker for their whole duration (Figure 5).

The selected shots were finally sorted by the log-likelihood ratio computed during the speech feature detection task.

rank	system	A.P.	D.100	D.1000	rank	system	A.P.	D.100	D.1000
1	**CL-LIMSI**	**0.721**	**100**	**997**	8	B_T1_1	0.645	95	934
2	B_M-1_1	0.713	99	990	9	B_T2_2	0.645	95	934
3	B_E2002_1	0.710	100	987	10	B_Sys1_1	0.645	97	932
4	B_l1_1	0.681	96	970	11	B_r1_1	0.642	92	936
5	B_l2_2	0.681	96	970	12	A_r2_2	0.630	95	924
6	B_Sys2_2	0.663	98	951	13	B_RA_1	0.570	100	792
7	**CL-GEOD**	**0.649**	**98**	**924**					

Table 3: Average precision and average hits at depth 100 and 1000 for feature 8 (speech).

rank	system	A.P.	D.100	D.1000	rank	system	A.P.	D.100	D.1000
1	B_M-1_1	0.268	14	37	6	B_l2_2	0.009	1	1
2	**CL-LIMSI**	**0.149**	**23**	**23**	7	B_RA_1	0.009	0	16
3	**CL-GEOD**	**0.117**	**14**	**14**	8	B_Sys2_2	0.009	1	14
4	B_r1_1	0.082	13	16	9	B_Sys1_1	0.008	1	14
5	B_l1_1	0.009	1	1					

Table 4: Average precision and average hits at depth 100 and 1000 for feature 10 (monologue).

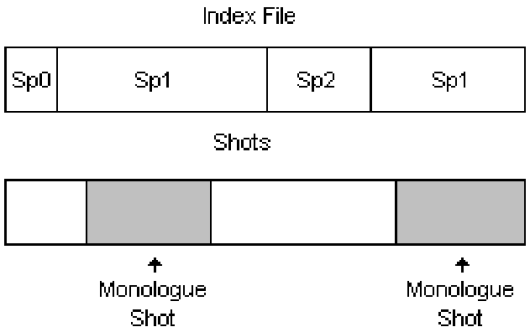

Figure 5: Monologue Feature Detection using Speaker Segmentation index file

3.2.3 Speech and Monolgue Features Evaluation

Alternatively to the above described system, we also used the output of the LIMSI Audio-Video transcription system [8]. This system is the one used for the LIMSI donated transcription for which we additionally had a speaker segmentation. The ranking was done using the same principles.

Table 3 and 4 show the performance of CLIPS-LIMSI and CLIPS-GEOD (described in sections 3.2.1 and 3.2.2) systems among other systems that have searched for features 8 and 10. The quality is very good for all systems for speech detection. LIMSI is ranked first and GEOD is in the average. The monologue detection is more selective and CLIPS-LIMSI and CLIPS-GEOD are ranked respectively 2 and 3 probably due to a good face detection.

4 Search Task

CLIPS-IMAG submitted three runs for the search task. One is based only on speech transcription (from LIMSI-CNRS), one based only on a combination of donated features, and one based on a combination of both. We did not use anything else like image similarity for instance.

A vectorial model was used both for the keyword-based search, for the combination of donated features, and for the combination of keywords and features. A weight can be given independently to each keyword (stemming was used) and to each donated feature. Independently weight can be given to the keyword based search and to the feature based search. A single system is used for the three runs. For the "ASR only" , the "ASR+features", and the "features only" runs, the keywords/features weights are respectively set to (1,0), (0.5,0.5) and (0,1). The selected keywords and features as well as their relative weight are chosen manually and once for the three runs.

Our three runs were manual only and of type A. However, the only use that we have made of the test corpus is an evaluation of the quality of the donated features (all of type B) in order to weight them accordingly. There is a fixed weighting of the donators for each feature according to a quality evaluation (which is combined to the weight of the features and to the keywords/features weights). Since the feature quality

rank	system	A.P.	D.10	D.100	rank	system	A.P.	D.10	D.100
1	M_B_ci_1	0.231	6.360	10.880	12	M_B_MT1_2	0.034	1.520	3.560
2	M_B_M-2_2	0.136	2.720	10.240	13	M_B_Aqt_3	0.026	0.480	3.600
3	M_B_UAL1_1	0.112	2.440	9.200	14	M_A_UAL2_4	0.026	0.320	4.920
4	M_B_M-3_3	0.093	2.240	9.160	15	M_B_MT2_3	0.019	0.880	2.280
5	M_B_0_T_2	0.092	1.920	7.240	16	M_B_eo.3_1	0.010	1.000	2.400
6	**CLIPS-ASR**	**0.071**	**1.560**	**7.240**	17	M_B_M-1_1	0.006	0.400	2.560
7	**CLIPS-A+F**	**0.064**	**1.520**	**3.840**	18	M_B_0_TIscG_4	0.004	0.120	1.040
8	M_B_KM-2_2	0.060	1.280	5.520	**19**	**CLIPS-Feat.**	**0.003**	**0.240**	**1.600**
9	M_B_qtrec_2	0.059	1.520	6.840	20	M_B_0_TIsc_3	0.002	0.080	1.400
10	M_B_KM-4_4	0.057	1.720	5.280	21	M_B_0_TIac_1	0.002	0.040	1.200
11	M_B_KM-3_3	0.043	1.160	5.320					

Table 5: Average precision and average hits at depth 10 and 100 for systems ran manually for the search task.

evaluation is the only use that we have made of the test corpus ans since we do not expect this quality evaluation to be very sensitive to this, our runs are almost of type B runs and we consider that the comparison with type B runs is meaningful.

Table 5 shows the performance of CLIPS systems among other systems that have processed manually all the 25 topics. Our "ASR only" and "ASR+features" runs ranked respectively 6 and 7 (on average precision) while the "features only" run ranked 19. Even though the topics were chosen in order not to favour speech recognition, the "ASR only" system performed slightly better than the "ASR+features" system. The feature only result is very poor probably because for many topics they are not very discriminative or even relevant.

5 Conclusion

We have presented the participation of the CLIPS-IMAG laboratory to the video track of the TREC-11 evaluation. We participated in all of the three proposed tasks. This participation was done in collaboration with teams from other institutions including LIMSI-CNRS (Orsay, France) for speech transcription, LIT-IPAL (Singapore) for face detection and INSA (Lyon, France) for text transcription. Our performance was quite good in shot boundary detection, average or poor for face and people detection, good for speech and monologue detection and quite good for the search task with speech recognition and poor without it.

References

[1] 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.

[2] 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.

[3] 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.

[4] Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 20, number 1, pages 23-38, January 1998.

[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] 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.

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

[8] Barras, C., Allauzen, A., Lamel, L., and Gauvain, JL,: Transcribing Audio-Video Archives. In *Proceedings of ICASSP*, pages 13-16, Orlando, May 2002.