

THE ARGOS CAMPAIGN: EVALUATION OF VIDEO ANALYSIS TOOLS

Philippe Joly^a, Jenny Benois-Pineau^b, Ewa Kijak^c, Georges Quénot^d,

^aIRIT, Toulouse, France

^bLABRI, Bordeaux, France

^cIRISA, Rennes, France

^dLIG-CNRS, France

ABSTRACT

The benchmarking of various methods of video analysis and indexing has become a research problem per se. The Argos evaluation campaign supported by the French Techno-Vision program aimed at developing resources for a benchmarking of video content analysis and indexing methods. The paper describes the type of the evaluated tasks, the way the content set has been produced, metrics and tools developed for the evaluations and some of the results obtained at the end of the first phase. Perspectives based on current works will be given to conclude this paper.

1. INTRODUCTION

Research in video analysis and indexing has a very wide range of applications in the overall activity of human society. This is why it has been very much competitive since the last decade. Taking into account numerous contributions of a large amount of research groups all over the world, the benchmarking and evaluation of different methods has become a must in this field. In this paper we present French national initiative, which is an evaluation campaign ARGOS supported by Technovision framework program.

The goal of this campaign being the evaluation of video content analysis tools, we identified a subset of the most common tasks in this wide domain. This set was divided in two parts. The first one, which was the object of the first evaluation phase, gathers mainly tasks aiming at extracting low-level features. The second part, which will be evaluated in a further step, is composed of tasks requiring results of the first part to extract complementary pieces of information from the content. In that way, we firstly intend to evaluate the impact of low-level-features extraction quality on higher-level feature analysis. The second original aspect of this campaign consists in its specific corpus design. The latter is composed of heterogeneous content entities thus limiting over fitting effects which can be observed when tools are evaluated only on one specific type of content

coming from only one source. Therefore, as it will be shown in the last part of this paper, results are generally lower than the one which can be expected on homogenous content sets.

Due to this heterogeneous aspect of the video data, the evaluated tasks are not as specific as they can be in campaigns which are more dedicated on the analysis evaluation of specific contents. Let us mention here for example the Trec Video campaign which has been up to now mainly focusing on TV News program indexing [1], the Etiseo campaign organized to evaluate analysis tools of video surveillance streams [2] or the Clear campaign aiming at the evaluation of indexing tools applied to meeting recordings [3].

As a general framework, the ground truth annotation format was mapped on the one used in Trec Video. To ensure a simple mechanism for the annotation and for the production of results, nearly all the tasks are specified as a verification process. To do so, all the documents of the content set have been initially segmented. Being given a set of segments, a task consists generally in verifying that a given feature (a specific camera motion for example) can be observed (or not) in each of them.

Results analysis was made regarding the type of the content set subpart, and the instance of the feature to be detected. The paper is organized as follows.

In section 2, we present the tasks proposed by the campaign. We then explain, in section 3, how the corpus has been gathered and annotated. In section 4, we introduce the metrics used for the evaluation, and finally, we present some of the results obtained after the first phase of the campaign.

2. EVALUATION TASKS

Evaluation tasks are related to basic topics gathering a large number of scientific contributions. They are separated in two groups corresponding to two successive evaluation steps in the campaign.

The first one leads to the evaluation of low-level feature extracting tools such as shot segmentation and transition

effect identification, camera motion identification, shooting location identification (inside / outside), person presence detection, and text presence detection.

The tasks of the second group are partially or fully based on the aggregation or the fusion of the first group results. Therefore, tools of this group may use the results of the first evaluation step. Furthermore, all tasks of the first group will be evaluated again by the end of the second step in order to evaluate mechanisms of reinforcement. The tasks of this second group are: framing classification, identification of the different appearances of a same person, ascii transcription of text in video, segmentation into stories, and human behaviour identification.

For the annotation and the evaluation mechanisms, we consider that there are two types of tasks:

- two segmentation tasks are evaluated regarding the ability of an automatic tool to precisely localize boundaries along the time axis. The first task is the segmentation into shots, the second one is the “story segmentation” already addressed by the Trec Video campaign [1].
- all the other tasks are aiming at automatically extract a subset of predefined segments where a given feature is present. As an example, here is the set of features evaluated for the tasks on shot transition and camera motion identification :
 - o Shot transition: cut, gradual interpolating transition (fade, cross-dissolve), gradual overlaying transition (wipe, etc), partial transition (compositing effect)
 - o Camera motion: still shot, zoom/dolly in, zoom/dolly out, pan/dolly to the left, pan/dolly to the right, tilt/dolly up, tilt/dolly down.

3. CORPORA AND ANNOTATION

The evaluation of a large variety of media indexing methods requires publicly available corpora of data cleared of IPR and privacy rights.

While the availability of broadcasted and archived artistic corpora is a matter of an IPR, the availability of realistic video data for video surveillance applications is the matter of privacy as well. Each figurant appearing in a video segment, according to the legacy regulations of the country has to explicitly give his approval for the use of his image for the purpose of this scientific work. This is why publicly available corpora for video surveillance applications are not very frequent. We can mention here the corpus of CAVIAR IST EU-funded project, which can be downloaded at [4].

As mentioned before, the content set of ARGOS campaign was constituted to address the goal of benchmarking video content analysis tools for specified

tasks on heterogeneous sources of content, but also the goal of benchmarking tools on contents registered in heterogeneous formats.

Thus, three sources of video content were registered: TV news journals, documentaries and video surveillance scenes. TV news journals were supplied by INA ®, which is a major audiovisual archive center in the world and the first digital image databank in Europe. The documentaries were supplied by SFRS-CERIMES®, which is the French center of multimedia information resources for Higher Education. Video surveillance content was produced by the members of ARGOS steering board: the IRIT and LABRI research centers in computer science. It comprises a rich variety of scenes registered in interior and exterior environment.

In order to test the robustness of the methods with regard to the quality of video source, the content was encoded in compressed formats MPEG1 and MPEG2 for both broadcast/artistic content and video surveillance content. Furthermore, for video surveillance content, both a professional CCD color video surveillance camera and a camera on IP were used to record the scenes.

The content of the corpus is distributed according to the Table 1.

Producer	Content type	Duration (min)	Number of files	Format	Resolution	Bit rate (MB/s)
INA	TV News	600	17	Mpg1	352 x 288	1
SFRS	Docs	624	21	Mpg2	720 x 576	6
LABRI	Video surveillance	632	13	Mpg2	352 x 288	2.7
IRIT	Video surveillance	600	20	Mpg1	352 x 288	1.3

Table 1. Content distribution in the corpus

Key frames of content items for video surveillance content are given in Figure 1.

The annotation of the corpus has been done accordingly to the source of content and the tasks the content has to be used in. Thus for video surveillance corpus, such annotation as shot boundary or camera motion would not make sense as the shooting was realized with a single camera (no shot boundary) which was static (no camera motion to be detected). On the contrary, the tasks of human behavior characterization had to be fulfilled on this corpus. Therefore the annotation of the video surveillance content set was done on consecutive segments of 20 sec duration.

The annotation interface for video surveillance corpus is shown in Figure 2.a. On this interface, the frame number as well as the segment reference, is permanently displayed.

The annotation of documentaries and TV news journals was fulfilled accordingly to the production rules. The classical task of shot boundary detection and transition effect characterization was easy-to-do due to the visualization of a buffer of video frames. The annotation of stories, according to a task of the second phase, supposed the identification of a linear structure of a video document as a set of consecutive groups of shots. Manual text transcript of titles and of other text appearing in a video frame was realized on TV news by the INA partner.

The annotation of camera motion was ensured by this interface as well. Nevertheless, it should be stated that the annotation of global motion in video is a tedious task. In TREC Video 2005, the annotated test set for camera motion characterization was very strongly filtered. The reason for this is that in case of hand carried camera or engine carried camera (boat, vehicle, airplane), the motion is never “pure” e.g. tilt or zoom. The visual interpretation is rather subjective. This is why in TREC Video 2005 only segments with a clear pre-dominant motion were used in test data set.

In ARGOS, all shots were annotated. For each feature, the annotators have specified whether the segment contained or not the searched feature. This annotation is exhaustive, unlike in the TREC case where pooling is generally involved. Additionally, the annotators have marked some segments as ambiguous or especially difficult using a “joker” tag, stating so the difficulty of the analysis or the possible lack of relevance of the annotation locally.

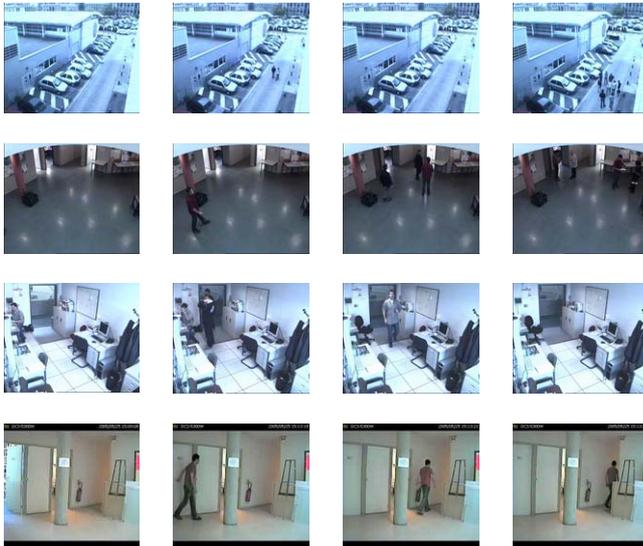


Figure 1: Excerpts of video surveillance content: 1st row: exterior – 2nd row Interior wide angle camera lens – 3rd row: Interior, closer caption, 4th row: Interior, lower resolution.

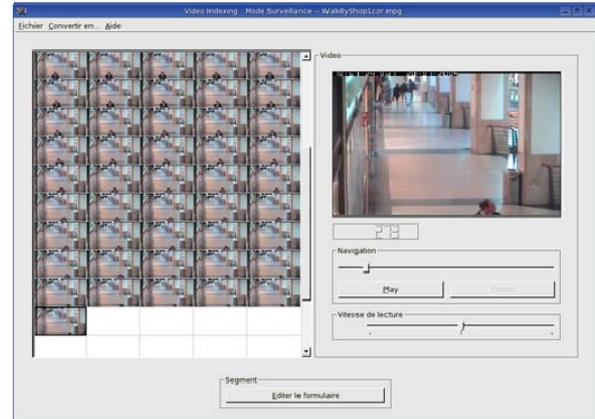


Figure 2: Annotation interface of video surveillance corpus

As far as several persons participated to the annotation, a set of rules was established in order to ensure the ground truth consistency. Nevertheless, the experience showed that humans rather subjectively characterize dominant motion components and thus the results of benchmarking on non-filtered reference annotation of this task, evaluated against the human perception of that concept of dominant motion, are not too high. Some examples of annotated motions are given in Figure 3 below on feature documentaries. Here the human operator annotated a “pure” motion while the apparent motion in each case is a combination of several components.



Figure 3: SFRS-CERIMES documentaries. Task on camera motion characterization: 1st row: zoom out – 2nd row: pan right – 3rd row: pan right – 4th row: zoom in.

Increasing the accuracy in the annotation process, to take into account possible variations of the interpretation of the content due to each person sensibility, could be possible while duplicating the process on different sites. Fusion of

results of each independent annotation would be of much profit for such tasks, but of course of a higher cost.

4. METRICS AND PROTOCOL

As mentioned before, Argos tasks have been split into two types: segmentation tasks and detection tasks. The segmentation tasks include shot segmentation and story segmentation. The detection tasks include the search for a variety of unrelated features including camera motion identification, person detection and event recognition. Both segmentation tasks are evaluated using the same metrics and protocol. Similarly, all detection tasks are evaluated using the same metrics and protocol regardless of any specificity in the detected feature.

4.1 Segmentation tasks

The segmentation tasks are evaluated using Precision and Recall based metrics. Two types of metrics have been considered. The first one, called “ARGOS” is based on a maximum overlap between extracted segments. The second one, called “TREC” is based on precision and recall over the detected transitions between video segments; it is similar to the one used in the Shot Boundary Detection (SBD) TRECVID task [1]. The F-measure, which is the harmonic mean of Precision and Recall, is also considered as a global performance indicator.

4.2 ARGOS metrics

A “reference segmentation” and a “system segmentation” are considered by the Argos metrics. The former is the one that has been manually generated by the annotators of the corpora. The latter is the output of a competing system. Both segmentations are represented by video segments corresponding either to shots or to stories depending upon the considered segmentation task. These video segments can be separated by “gaps”; for instance, in the case of shot segmentation, the gaps correspond to the duration of progressive transitions (e.g. dissolves or wipes) which are not considered as part of the shots. Gaps can occur both in the reference segmentation and in the system segmentation.

From both reference and system segmentations, a “maximum intersection” segmentation is defined by (see Figure 4): intersection segments must be completely included in a reference segment and in a system segment; a system segment or a reference segment can include at most one intersection segment; the sum of the lengths of the intersection segments is maximized. Finding the maximum intersection involves finding an optimal matching between reference and system segments. The matching is optimal in the sense that the length of the associated intersection is maximized. Such a matching is generally unique but not always. In the case it is not unique however, all the matching leads to the same intersection length. The lengths of reference, system and intersection segments are defined

as the number of frame they contain. In practice, the maximum intersection and the maximum length are computed using Dynamic Programming (DP). DP allows virtually trying all the possible matching and selecting the maximum intersection in quadratic time relatively to the number of segments.



Figure 4: Definition of the maximum intersection from reference and system segmentations.

Once the maximum intersection is defined, ARGOS Precision and Recall measures are defined as:

$$\text{ARGOS_Precision} = \frac{|\text{intersection}|}{|\text{system}|}$$

$$\text{ARGOS_Recall} = \frac{|\text{intersection}|}{|\text{reference}|}$$

$$\text{ARGOS_F-measure} = 2 \times \frac{|\text{intersection}|}{(|\text{reference}| + |\text{system}|)}$$

where $|\cdot|$ represent the sum of the lengths of the segments of the considered segmentation (excluding gaps).

4.3 TREC metrics

The TREC metrics is based on Precision and Recall on detected transitions. The reference and system outputs are transformed into a list of transitions, either “gradual” or “cut” depending whether there is or not a gap between the video segments. An optimal matching between reference and system transitions is searched for: transitions can be matched if they overlap each other; a system transition can match at most one reference transition and vice versa; the number of matches is maximized. Again, DP is used to find an optimal matching. Once the optimal matching is found, TREC Precision, Recall and F-Measure are defined as:

$$\text{TREC_Precision} = \frac{|\text{matched}|}{|\text{system}|}$$

$$\text{TREC_Recall} = \frac{|\text{matched}|}{|\text{reference}|}$$

$$\text{TREC_F-measure} = 2 \times \frac{|\text{matched}|}{(|\text{reference}| + |\text{system}|)}$$

where $|\cdot|$ represent the number of transitions in the considered set.

4.4 Comparison of the metrics

The TREC metrics has been introduced to ease the comparisons with TRECVID evaluations. The ARGOS metrics has been chosen because it is less sensitive to some effects that are not very relevant for the possible applications. First, it is much less sensitive to approximate boundaries. If a user wants to visualize the content of a video segment, a few extra or missing frames at the beginning or at the end will make a little difference. This tolerance avoids penalizing methods which work in the compressed domain only on I-frames or P-frames. These methods can be much faster and of comparable quality. Second, the ARGOS measure is less sensitive to over-segmentation in highly changing segments. The truncation of small segments is expected to be less penalizing in

applications than merge and split of larger parts of segments.

In practice, the ARGOS Precision and Recall are much more correlated and close to each other than the TREC ones. This is due to the fact that the gaps often represent a negligible fraction of the total length of the video documents. Therefore, for the ARGOS metrics, $|\text{system}| \approx |\text{reference}| \approx |\text{video}|$ and then $\text{Precision} \approx \text{Recall} \approx \text{F-measure}$. Also, when a parameter is changed into the competing system to adjust the degree of segmentation from under-segmented to over-segmented, TREC_Precision continuously decreases and TREC_Recall continuously increases while ARGOS_Precision, ARGOS_Recall and ARGOS_F-measure change in a correlated way by first increasing up to a maximum value and then decreasing. From this respect, there is a natural optimum value for the tuning parameter according to the ARGOS measure. The TREC_F-measure behaves in a way similar to ARGOS measures.

4.5 Detection tasks

The detection tasks are evaluated using Precision and Recall based metrics. A reference segmentation is given for the whole corpus. For TV documents (INA and SFRS collections), the reference segmentation is the segmentation into shots; for video surveillance documents (IRIT and LABRI collections), the reference segmentation are segments of fixed duration of 20 sec. The detection tasks consist in deciding for each video segment whether some features are absent or present. The evaluation is done at the level of each feature to be detected and also at the level of the task (taking into account all the features in this task).

Let us denote R the set of shots in which the feature has been judged present by human annotator and J , the set of shots judged as especially difficult (joker tag, see section 3).

Let us denote S the set of shots the competing system detected the feature in. We will then define Precision, Recall and F-measure as:

$$\text{Precision} = |S \cap R| / |S|$$

$$\text{Recall} = |S \cap R| / |R|$$

$$\text{F-measure} = 2 \times |S \cap R| / (|R| + |S|)$$

where $|\cdot|$ denotes the number of elements (shots) in the set. We will also define “relaxed” Precision, Recall and F-measure as:

$$\text{Relaxed_Precision} = |S \cap R \cap \neg J| / |S \cap \neg J|$$

$$\text{Relaxed_Recall} = |S \cap R \cap \neg J| / |R \cap \neg J|$$

$$\text{Relaxed_F-measure} = 2 \times |S \cap R \cap \neg J| / (|R \cap \neg J| + |S \cap \neg J|)$$

where $\neg J$ denotes the complementary set of J in the set of all shots (excluding ambiguous or too difficult cases).

4.6 Evaluation Tools

Tools have been developed for the automatic assessment of the system performance. These tools can be downloaded from the ARGOS site [5] in the section evaluation. They can be used by the ARGOS participants to check their results or to train their systems using the data and reference

annotations from the previous phases. They can also be used for other evaluation campaigns provided that the reference and system file format are respected. These tools take as input one system submission file and one or two reference files. They automatically compute the above described metrics at various levels of details. Results are given by default on the whole ARGOS corpus but they can also be given at the collection level (e.g. SFRS) or at the video document level (e.g. SFRS12). Finally, lists of false positive, false negative and correct shots can be given for fine grain analysis (for instance in order to make statistics to identify especially difficult cases). For the segmentation tasks, results can be computed either with the ARGOS metrics or with the TREC metrics. For the detection tasks, results can be computed either on all shots or on all but the ambiguous or too difficult ones.

5. RESULTS OF THE FIRST PHASE

Submission of several runs was allowed for participating groups, which could determine different parameter settings for each submitted run. To illustrate results of the campaign, we focus on one segmentation task and one detection task among the six tasks proposed in the first phase. The shot segmentation and camera motion detection results are presented below.

5.1 Shot Segmentation task

Five groups submitted a total of twenty-two runs for the shot segmentation task. The approaches are first briefly presented. The approach taken by Eurecom is based on frame similarities measured on region-based HSV color histograms. Cuts and gradual transitions are detected in one-pass, using separate decision methods. CLIPS-IMAG detects cuts by image comparisons after motion compensation. Gradual transitions are detected by comparing norms of the first and second temporal derivatives of the image. IRIT computes intensity gradient for each RGB component. The threshold used for cut detection depends on activity rate. The sign of second derivative is considered to discriminate gradual transitions. LaBRI uses approach in the compressed domain. Shot detection is based on 3 measures: camera motion continuity computed from MPEG motion vectors, frame statistics as number of intra-coded macroblocks, and similarity between compensated adjacent I-frames. ViperUnige approaches the task using interest points descriptors extracted every 10 frames. Interest points are then matched and a matching distance between frames is computed based on SIFT descriptors. A threshold is applied on this distance to detect cuts.

Precision and recall measures using ARGOS and TREC metrics are compared in Figures 5 and 6. They illustrate the metrics variations in the range of under to over segmentation. As mentioned before, precision and recall are highly correlated in ARGOS metrics. Under-segmentation

and over-segmentation are both characterized by a small ARGOS_Precision and Recall, while under-segmentation (respectively over-segmentation) is characterized by a small (resp. high) TREC_Recall and a high (resp. small) TREC_Precision. Some methods have their results highly decreased by the TREC metric. As TREC metric is sensitive to transition location, it happens that the global segmentation is good but transitions are systematically misplaced.

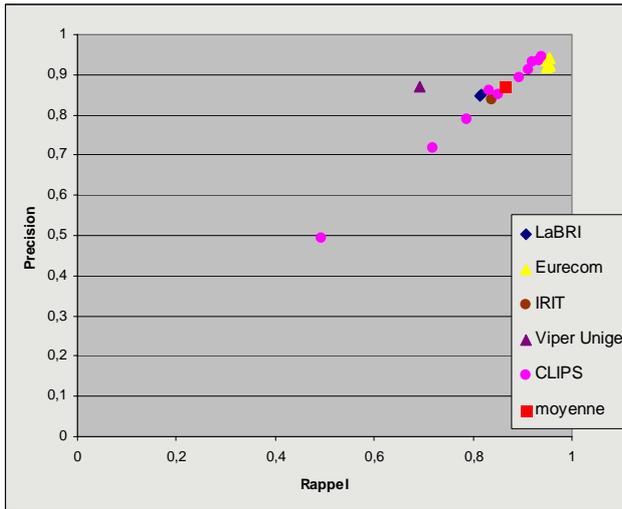


Figure 5: Precision and recall for shot segmentation using ARGOS metrics.

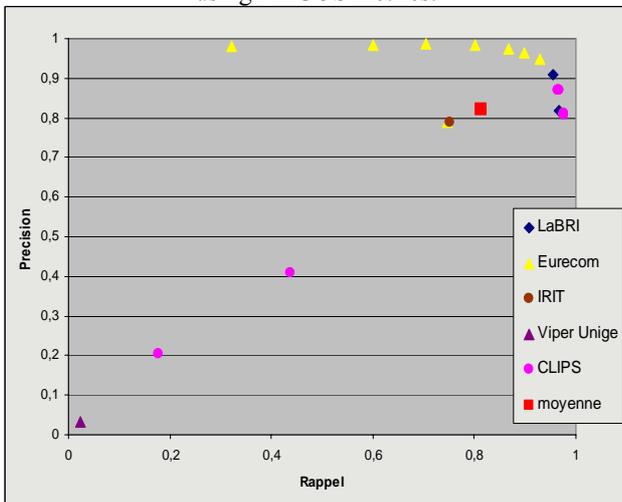


Figure 6: Precision and recall for shot segmentation using TREC metrics.

Two corpora, TV news and scientific documentaries, were used for shot segmentation evaluation. Cuts represent about 95% of all transitions (see Figure 8). Gradual transitions are represented by gaps between video segments in the shot segmentation task. The results separately produced on each part of the corpus were compared. Some

differences between these results can be observed. According to the results of transition effect classification task (which is not discussed here), gradual transition identification is better performed on the TV news corpus. Figure 8 highlights this phenomenon. The documentaries part of the content set is the most heterogeneous one compared with the complete collection. Therefore, adapting indexing tools on this part is more complex than for the TV News collection.

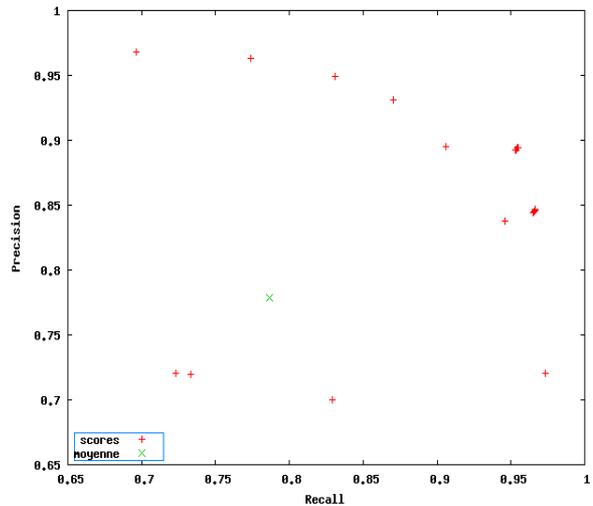
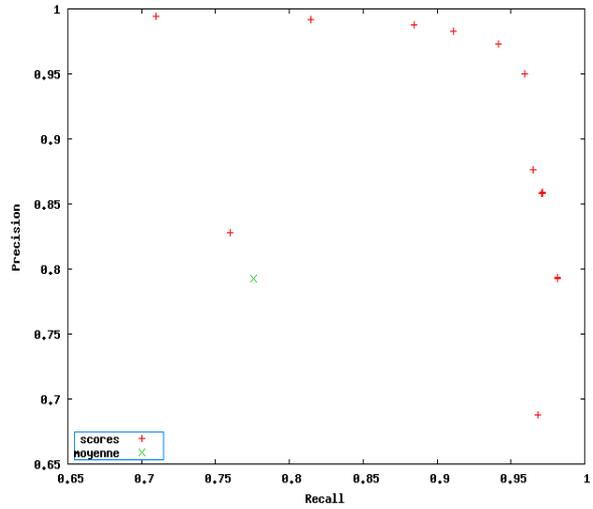


Figure 7: mean precision and recall for transition detection on the TV-News (for the first picture) and on the documentaries (for the second one) sub-part of the content set.

Results for each video of the corpora were also compared to point potential differences between videos. Except one video for which results were lower than average, no significant differences were identified. Finally, shots that present difficulties whatever the system (and parameters) employed were highlighted. The harder transitions to detect are partial transitions or cuts resulting from removal of part

of a scene. Presence of flash inside a shot is the main source of false detections.

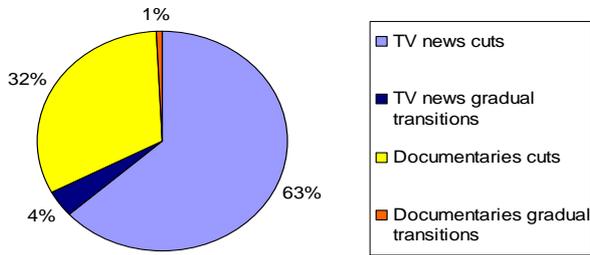


Figure 8: Distribution of transition types on TV news and scientific documentaries corpora.

5.2 Camera motion identification

Four groups submitted a total of twenty runs for this task. Six features corresponding to three groups are defined: zoom in and out, pan left and right, and tilt up and down. The distribution of features on the corpus is given in Figure 9. Near 30% of the considered shots are tagged with a “joker” as explained above.

Participants’ approaches were varied. LaBRI used MPEG motion vectors to build a 6 parameter affine model. A statistical significance test was performed to identify elementary motion. Eurecom estimated a 4 parameter affine model using least square minimization on optical flow. ViperUnige employed trajectories of interest points to estimate the affine motion. Classification was performed using thresholds. IRIT computed spatiotemporal X-RAY and Y-RAY images. These images were quantified and segmented into regions. Then an edge orientation histogram was computed and motion was identified by the max value of the histogram.

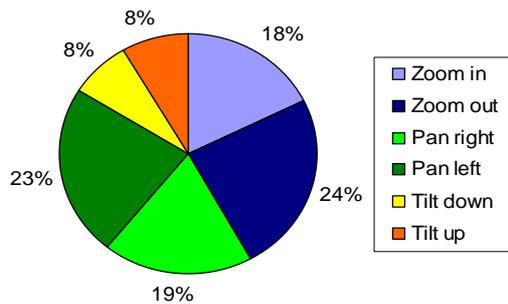


Figure 9: Feature distribution over the corpus.

Information on results is depicted in Figure 10. Participants obtain higher results for zoom in, followed by zoom out then pan and tilt. Using MPEG motion vectors and robust estimators gives higher results than using optical flow. It seems that spatiotemporal images are more efficient for right/left pan classification than up/down tilt.

Figure 11 shows “relaxed” precision and recall (as defined in section 4.2) for some groups of camera motion.

This points the general trend of different methods between under-detection (statistical tests) and over-detection (threshold).

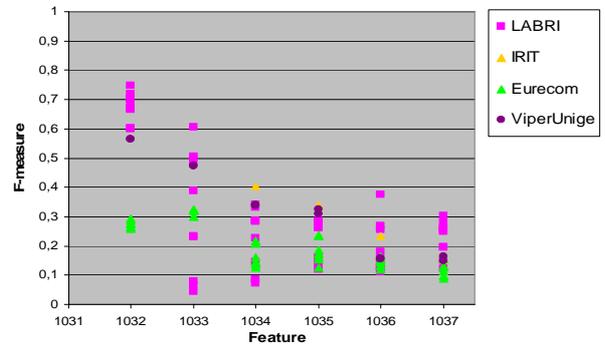
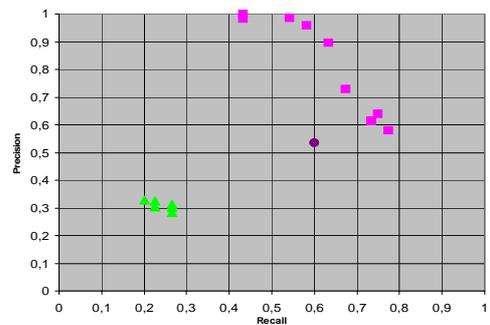
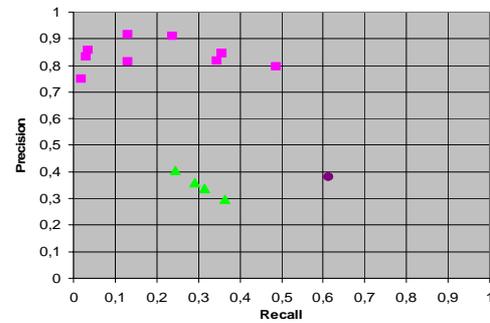


Figure 10: F-measure for zoom in (1032), zoom out (1033), pan right (1034), pan left (1035), tilt down (1036) and tilt up (1037).

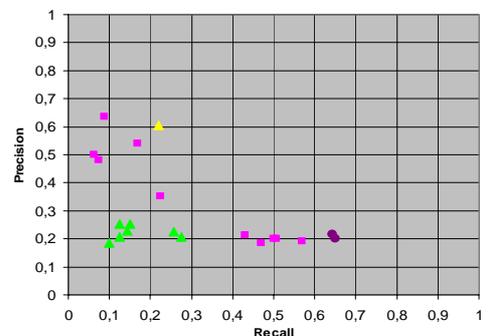
Zoom in



Zoom out



Pan left



Tilt
down

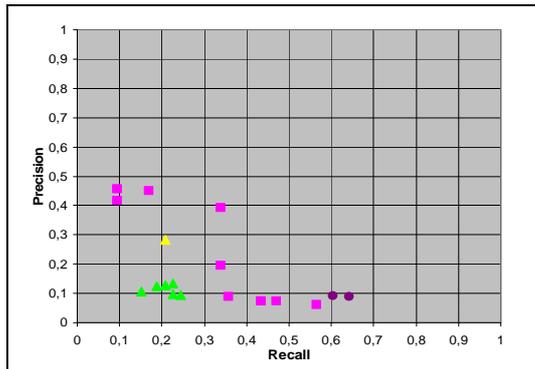


Figure 11: Precision and recall for different camera motion classes.

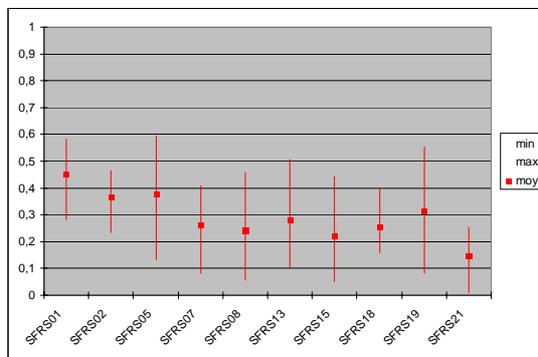


Figure 12: F-measure mean values (red squares) and variations of camera motion classification tools obtained on different video files of the documentaries collection.

Figure 12 confirms that the performance between the analysis tools is significantly varying. But we can also observe that this performance is also significantly varying between the documents (from 25 to 60% for the best results).

6. CONCLUSION AND FURTHER WORKS

As an immediate valorization of that work, some of the developments will be used in different consortia such as Cost 292 and the InfoMagic project.

During the meeting sessions of the campaign, researchers mentioned their interest for an evaluation process allowing to compare performances of different algorithms more or less fitted to a given sub part of the content set. In that sense, the evaluation software and the ground truth are expected to be interesting evaluation tools outside the scope of the campaign. We are now considering the possibility to not distribute the ground truth by the end of the campaign, as it was planned at the earlier step of the project, but to open a permanent access to an online evaluation engine, on the Internet, which will allow to submit results on any task,

on any part of the content set, at any time, to be given results with any metrics. Of course, that kind of approach will definitely remove any interest in ranking the evaluated tools regarding to a given metric or another one. But it may become a reference able to associate a “score” to a tool, and to compare it with any other one evaluated before or later on the same site with the same data. This idea of an online evaluation tool will be also explored in two different forthcoming projects:

- in “Casewp” (which is a showcasing project developed in the framework of the “Muscle” Network of Excellence funded by the European Community), an original content set is currently produced and will be proposed to the research community with an online evaluation service dedicated for segmentation tools.
- In “Quaero” (a French project on digital multimedia content indexing), a large set of indexing tools will be evaluated on a large set of heterogeneous contents. We will explore the idea integrated in an “evaluation campaign management tool” which shall allow handling authorizations to submit results on the right contents at the right time, and to produce synthetic results in the shortest delay.

Today, being given an access to free annotated video data remains a problem. Specific corpora are usually developed by researchers to evaluate their own technology. Considering the human cost of this process, sharing these corpora, and more specifically the ground truth, could be appreciated outside the scope of an evaluation campaign, as a research tool *per se*.

7. ACKNOWLEDGEMENTS

The authors would like to thank the persons in charge of the Technovision Program, all the partners (CLIPS, Eurecom, INA, IRIT, LABRI, LIP6, NOVELTIS, and SFRS) who are involved in the campaign organization and all the participants.

8. REFERENCES

- [1] TRECVID 2005 An Overview P. Over, T. Ianeva, W. Kraaij and A. F. Smeaton, TRECVID'2005 Workshop, Gaithersburg, MD, USA, November 14-15, 2005.
- [2] <http://www.silogic.fr/etiseo>
- [3] <http://www.clear-evaluation.org/>
- [4] <http://groups.inf.ed.ac.uk/vision/CAVIAR/>
- [5] <http://www.irit.fr/argos>
- [6] J. Nesvadba, F. Ernst, J. Perhac, J. Benois-Pineau, L. Primaux, “Comparison of shot boundary detectors”, ICME'2005, pp788-791, Amsterdam, NL.