Shot Boundary Detection in the framework of Rough Indexing Paradigm

L. Primaux, J. Benois-Pineau, P. Krämer, J-P Domenger

LaBRI CNRS UMR 5800
351 cours de la Libération, Université Bordeaux 1
33405 Talence Cedex
France

Abstract

This paper presents the Shot Boundary Detection system developed by LaBRI in the context of "Rough Indexing" paradigm. We work on compressed streams and we use only I and P frames information, (DC coefficients of I-Frames, motion vectors of P-Frames and DC coefficients of prediction error) which allow us to be faster than many equivalent systems (10 times faster than real-time on TRECVID2003 test set, and 3 times faster on 2004, because MPEG files structure is composed of only I and P frames). In this context the application was not developed to classify shot change transition effects, the initial goal was to allow a real-time and intelligent browsing in video content for common users.
The detection is performed in two stages:

- Robust Global Camera Motion Estimation
- Detection of P-Frame peaks (computation of motion and frame statistics), and of I-Frames (measuring similarity on successive compensated I frames).

As we work with two types of frames (I and P), we associate two statistical models which give us two sets of ratio and threshold to calibrate the detector.
The first TRECVID participation of LaBRI implies an evolution of the application for transitions effects distinction, which induces two new thresholds to calibrate.
We generally obtain equivalent values of Recall and Precision (0.72 on TRECVID 2003 test set). On TRECVID 2004 test set we obtain as best runs ***ri-3***: 0.723(Recall) and 0.606(Precision); and ***ri-4***: 0.703(Recall) and 0.635(Precision).

1. Introduction

With the digitalization (MPEG) of old VHS archives and emerging intelligent home multimedia devices, it is necessary to develop fully automatic and fast segmentation algorithms. The first step in Scene (or Story) Segmentation is to detect Shot changes; the second is to group those shots into scenes. There are many ways to perform Shot Boundary Detection, such as the others participants approaches, which are developed in the following part.

1.1 Related Works

In the edition 2004 of TRECVid there were 19 attendees for the Shot Boundary Detection task. From the previous editions, most approaches worked in the

uncompressed video domain [1]. Those algorithms give relatively good results but in general they are slower than real-time.
Most methods are based on frame comparison (dissimilarity measure) such as pixel-by-pixel frame comparison [2], which gives good results but induces a very high complexity and it is not robust to noise and camera motion. As well as frame content representation by histograms and vector distances measures produce a good frame dissimilarity measure [3], but histograms lack of spatial information. This lack needs to be compensate with local histograms [4] or edge detection [5].
CLIPS system uses direct image comparison for cut detection [6]. In order to reduce over-detection, frames are compared together after motion compensation and a separate camera flash detection is also used. Gradual transitions are detected by checking if the pixel intensity follows a linear function along successive frames.
Systems based on histograms seem to obtain best results on recent TRECVID sessions. IBM proposes a system [7] which employs a combination of three-dimensional RGB colour histograms and local edge gradient histograms. Adaptive thresholds are computed by using recent frames as reference. MSR-Asia system [8] uses global histograms in the RGB colour space.
There are only few participants who work on compressed streams [9]. As they do not fully decompress the frames the complexity is reduced, this implies a loss of detection accuracy. The ATL [10] system uses a hybrid method to reduce complexity and keep detection accuracy. Only DC coefficients of I-Frames are used to perform the shot change localisation (by using histograms distances), then intermediate frames are decoded and compared to refine the SB detection. The KDDI system [11] works on partially decompressed streams by comparing DC coefficients of I, P and B frames, the results were promising.
Adaptive thresholds are essential to perform the detection, independently of the domain (compressed or uncompressed stream) and of the dissimilarity measure quality. It has been shown that fixed thresholds were ineffective on non-homogeneous video [12] such as news video content proposed by TRECVid.

1.2 TRECVid evaluations requirements

TRECVid proposes a Shot Boundary detection task, which is very much appropriated for professional applications. Not only the exact position of shot change is required but also the classification of shot transition effects in two categories: CUT and GRADUAL transition. Gradual transitions must be exactly delimited, which is a very complex task due to the variety of those transition types (fade, dissolve, slide, etc).

1.3 Rough Indexing paradigm

This framework, introduced in [13], is aiming to get an approximate solution with rough data. It imposes to extract as less information as possible from a video stream before any treatment. MPEG or H26x files can be partially decoded in order to get only I-Frames at DC resolution, P-Frames motion vectors and prediction error at DC resolution. We also extract the map of intracoded macroblocks for each P-Frame, while B-Frames are not taken into consideration.
From those extracted information, we firstly estimate Global Camera Motion, then we deduce the number of Intracoded/Outliers macroblocks. That will allow us to measure

motion and content continuity of adjacent P-Frames. Indeed, we assume that an increase of the number of intracoded macroblocks is closely linked to a significant variation of the content, such as a shot change or the appearing of a new object in the scene. In order to improve the robustness of our method, we also develop a new similarity measure on I-Frames [15].
Finally, in the field of our participation to TRECVid evaluation campaign, we recently improve the system with a transition classification module.

2. Frame similarity measures in the context of Rough Indexing Paradigm

The core of our approach is based on the Robust Global Camera Motion Estimator described in [14]. The main advantage of such an estimator is its non-sensitivity to moving objects in the frame. We will take advantage of that in evaluating the motion continuity as it is detailed in the following.

2.1 Measure on P-Frames

In this section, the way we estimate the similarity measure on P-Frames is explained.
First of all, the motion continuity (MC) is determined along successive P-Frames using the method presented in [14], which consists in evaluating the sum of all absolute normalized differences between the six parameters of two successive frames.
Then, the number of intracoded macroblocks (Q) is deduced from the map of intracoded macroblocks. In the TRECVid system version we use its derivative form (ΔQ).
Finally, we define a linear combination (see equation 1) of MC and ΔQ, which represents our similarity measure for P-Frames. High values imply a strong dissimilarity between frames, on the contrary of low values.

$$D = \Delta Q^{\beta} * MC^{1-\beta} \quad (1)$$

here, β is a parameter set to 0.8 by default.

2.2 Measures on I-Frames

We introduce a new dissimilarity measure by I-Frame mapping [15] in our Rough Indexing framework. This section presents a summary of this approach.
In order to superimpose one DC I-Frame to another to measure their dissimilarity, a complete motion trajectory has to be estimated between this two frames. We dispose of the Global Camera Motion model for each P-Frame, but it is not sufficient to calculate the global motion for the whole sequence between two I-Frames. Motion model is needed for I-Frames as they do not contain motion information. Therefore, the motion parameters of I-Frames are extrapolated by a weighted linear regression in each GoP [15 - 3.2]. Thus, the trajectory of a block in an I-Frame to its previous I-Frame is known. Using the scaling factor, this trajectory can be computed for pixels in the corresponding DC images of I-Frames. Moving object macroblocks are not taken into account for matching, they are given as outlier macroblocks by the Robust Global Motion Estimator [14]. We can finally compare those pixels for matching the two DC images of I-Frames. An illustration of this process is presented on figure1.

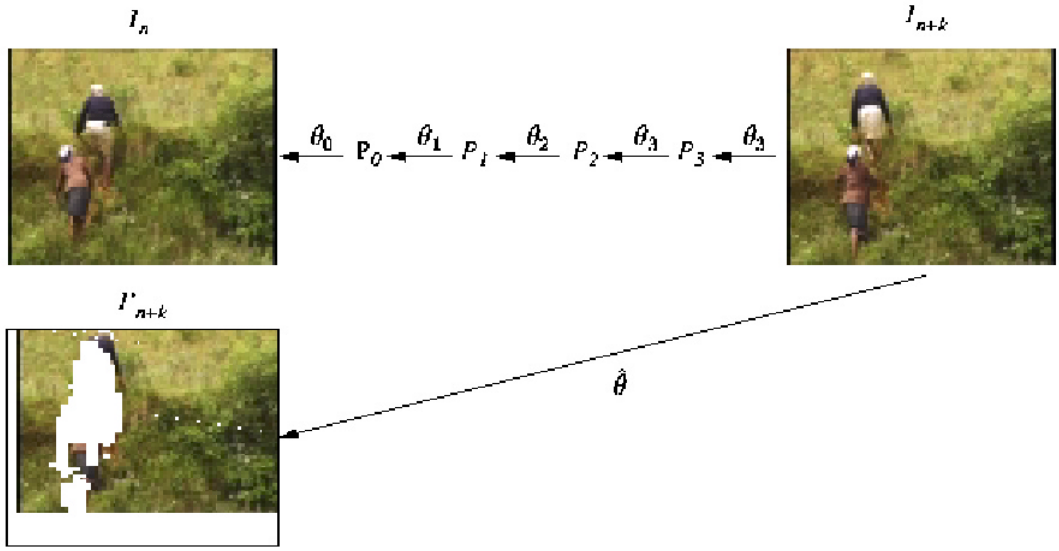


Figure 1 Projection of I-Frame I_{n+k} onto I-Frame I_n

here θ_i are camera motion models (6-parameters), we can see that the last θ needs to be extrapolated because I-Frame I_{n+k} has no motion information relatively to any frame. The first extrapolation method was to repeat θ_3, then we developed a weighted linear regression of the parameters along a GoP.

The matching of I-Frames takes into account luminance and chrominance components. The dissimilarity measure of I-Frames is a Weighted MSE calculated as [15 - 4.2]:

$$WMSE_{n+k} = \frac{1}{|V|}\sum_{p\in V} w(x'_p, y'_p)^2 * (DC_{n+k}(x_p, y_p) - DC'_n(x'_p, y'_p))^2 \qquad (2)$$

where $w(x'_p, y'_p)$ is a weight depending of a local contrast [15 - 4.2]. Thus it is adapted to the high frequency noise on DC frames. This measure is high if the content of two DC frames is different.

3. Shot Boundary Detection on I and P frames

3.1 Automata

The automata presented on Figure 3 shows only the shot change detection process, transition characterization will be described later. This method is an upgraded version of the method described in [14].

As it can be seen, P and I Frames peak detections are jointly fulfilled. In this way we can directly choose between a P-Frame detected shot change and an I-Frame detected shot change when they are close (less than one GoP size). A real P-Frame shot change often implies a high dissimilarity between the next I-Frame and the previous one, in this case the over detection on I-Frame is ignored. In the other case, when a P-Frame detected shot change just follows an I-Frame detected shot change, we only consider the I-Frame one.

Detection parameters appear in bold character in the automata, they will be described in the next section.

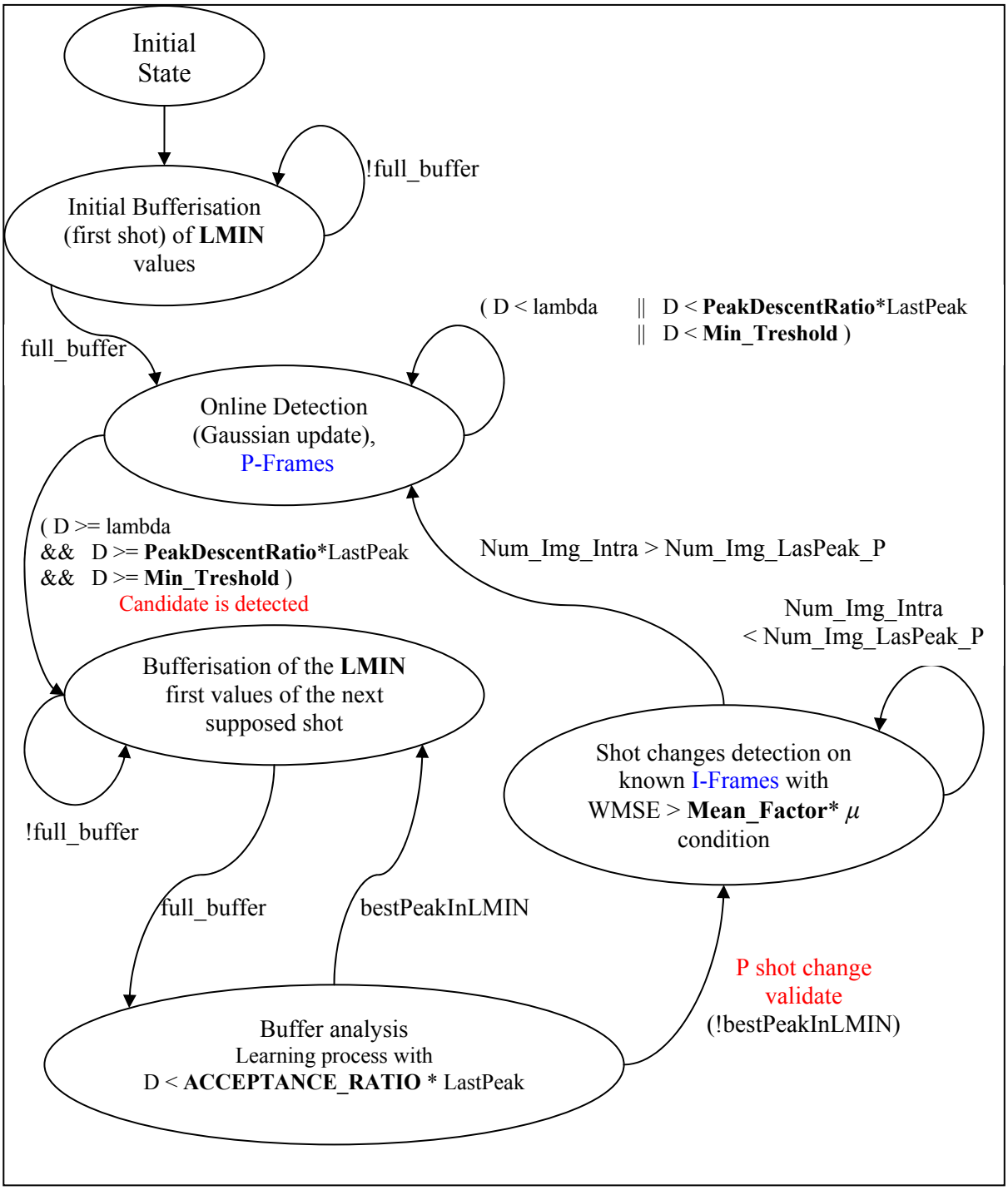


Figure 2 Shot Boundary Detector automata

3.2 Detection on P frames

The shot change detection is performed as follows (see figure 1). Shot transitions are detected as peak values of D along the time. For this purpose **LMin** minimal length of shot in terms of P or I frames is defined. In our case we take LMin=8. At the beginning of the process (i=0) or after each validated shot change detection (i=index of the shot change frame number), the mean μ and standard deviation σ are computed on the statistics D, in the interval L= [i+1, i+LMin] by using acceptable measures.

That means that we defined an **Acceptance_Ratio** parameter which allows us to remove from the learning process all values higher than *Acceptance_Ratio*D(i)*. We set this ratio to 0,05 by default. Supposing Gaussian distribution of D measures, following threshold value is defined:

$$\lambda = \mu + Kp * \sigma \tag{3}$$

where ***Kp*** is chosen from the interval [1 to 10] and in our case it is set to 1,8. Interval L represents frame from the beginning of new supposed shot and like this we try to estimate statistics of the measure D for frames in the new shot. Out of the L interval, μ and σ are updated as follows:

$$\begin{aligned} \mu &= (1-\alpha)\mu + \alpha\rho(\mu,\sigma,D(i))\mu \\ \sigma &= \sqrt{(1-\alpha)\sigma^2 + \alpha\rho(\mu,\sigma,D(i)) * (D(i)-\mu)^2} \end{aligned} \tag{4}$$

here $\rho(\mu,\sigma,D(i)) = 1$ in order to let Gaussian more reactive, and α is set to 0,15.

The index i is retained as shot border, if *D(i)* is a "peak outlier", that is:

$$D(i) > \lambda \tag{5}$$

But still, to recognize *D(i)* as a shot change peak the ratio between *D(i)* and previous shot change peak value *D(j)* has to be higher than the **PeakDescentRatio** predefined value:

$$D(i) / D(j) > PeakDescentRatio \tag{6}$$

This is done to avoid the acceptation of maxima which are not shot change peaks, under the supposition that successive shot changes should not differentiate very much in their order. In our case **PeakDescentRatio** is set to 0,15. We also add, based on TRECVid2003 test set experiments, a minimum threshold **Min_Threshold** which reinforce **PeakDescentRatio** false detection rejection capacity. This **Min_Threshold** is set to 25 on MPEG1 TRECVid test sets.

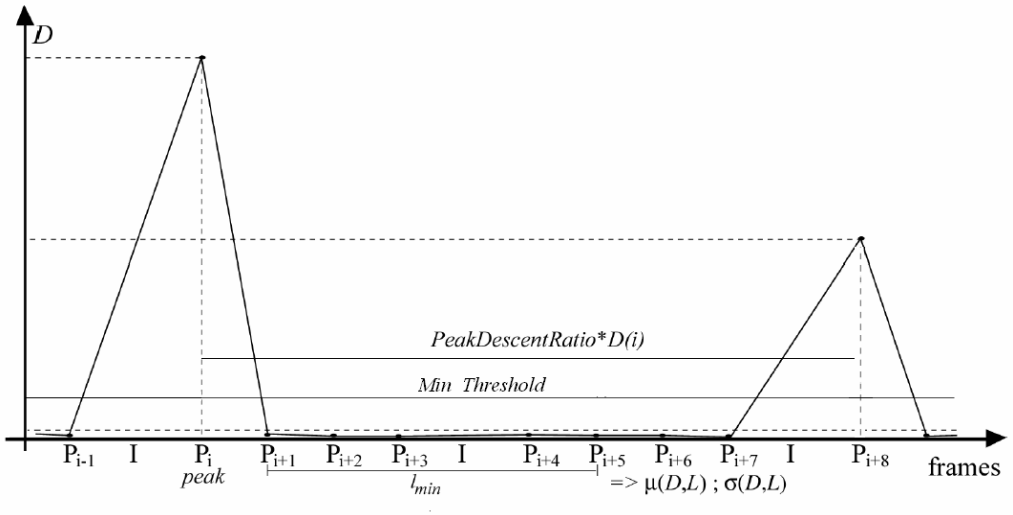


Figure 3 Peak detection in the characteristic D.

Finally, before to validate a peak as a correct shot change, we have to test if there is no best peak (higher value) in the LMin next frames.

After any P-Frame shot change has been validated, the I-Frame shot change detection module searches for shot borders since the previous P-Frame validated shot change. In the case of the first P-Frame validated shot change, the I-Frame shot border search starts from the first frame of the video.
The next section describes the I-Frame shot change detection method.

3.3 Detection on I frames

Shot change detection on I-Frames is performed by the following simple decision rules:
In each supposed shot we compute the mean μ_mse of Weighted MSE values (2). Then if the current Weighted MSE value is higher than *Mean_Factor** μ_mse the current I-Frame is considered as a shot change. **Mean_Factor** is set to 3, adapted by TRECVid 2003 test set experiments.

4. Shot transition effects classification

In this part our transition classification method is described. In the framework of Rough Indexing paradigm this detail level was not required. So the following method has been developed.

4.1 Transitions on P frames

The method described in [14] used the original number of macroblocks on P-Frames. We defined the notion of peak *density* and peak *width*.
Considering as D(i) the value of the current detected and validated peak, the *density* is the number of peaks which are higher than GRAD_RATIO*D(i) in the interval of 2*LMin frames [i-LMin, i+LMin]. The *width* is the number of successive higher values than GRAD_RATIO*D(i) which constitutes the current peak.
Then a "GRAD" was detected when both *width* and *density* are greater than a predefined threshold (chosen as 1). In fact, when considering progressive changes we observed the very chaotic behavior of motion parameters. Thus the absolute differences in the area of progressive changes and D(i) in consequence are high.

GRAD_RATIO has been set to 0,35 in order to get the most equilibrated results on test set of TRECVid2003.

We very recently upgraded the system by working on derivative of the number of intra-coded macroblocks. The classification consists now in considering as "CUT" all peaks which are immediately followed by their opposite, with the tolerance of one I-Frame in between. Moreover the peak must not be preceded by a high positive value. All other situations are considered as "GRAD". Figure 4 illustrates the two methods.
The decision based on the derivative of the number of intra-coded macroblocks showed an increased performance of 20% in recall. The corresponding results are given in section 5.

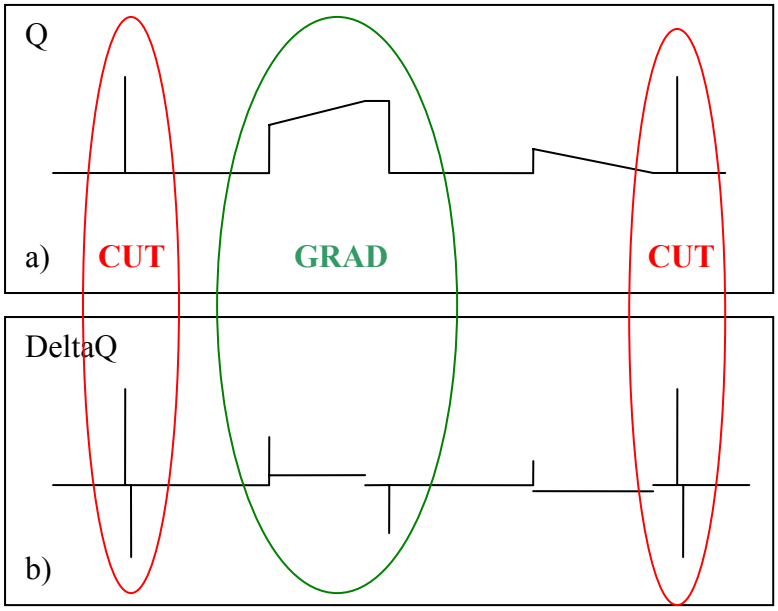


Figure 4 The two methods for transition classification, a) Decision with the value of intra-coded macroblocks number (Q); b) Decision with the derivative of Q

4.2 Transitions on I frames

Here we consider k-th value of dissimilarity measure (2) on I-Frames. The peak is classified as a gradual transition if the value of (2) for at least one of the two neighboring I-Frames is greater than *GRAD_RATIO_I* * *WMSE(k)*. It is clear, that such a rule is efficient for gradual transitions which are at least as long as a GoP.

GRAD_RATIO_I has been set to 0,4 for the same reason than for **GRAD_RATIO**.

5. Results on TRECVid2004 and Perspectives

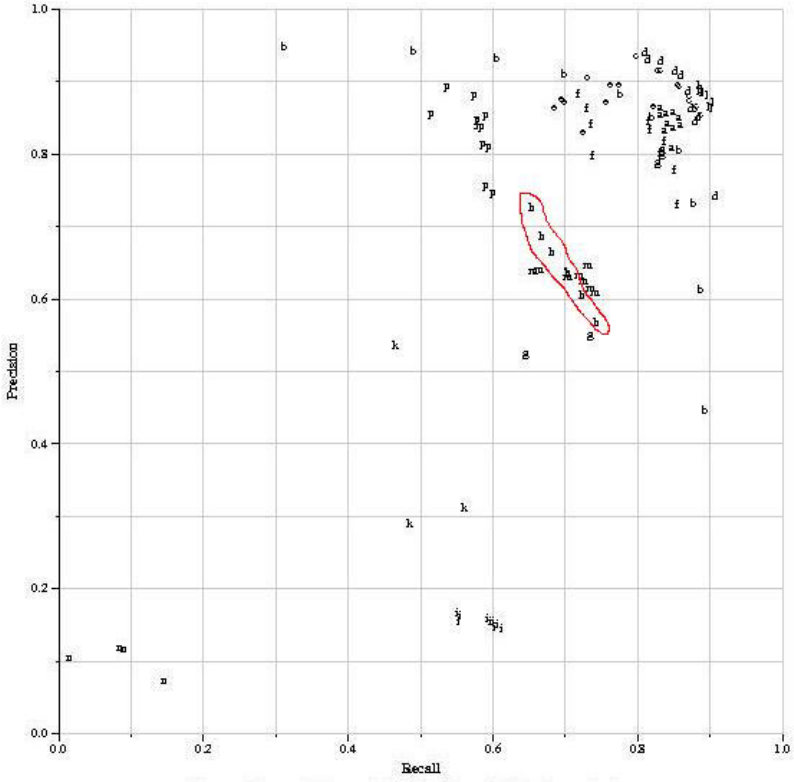
On TRECVid 2003 test set we manipulate 3 parameters in order to obtain different Recall/Precision compromises: Kp, Mean_Factor and β.
The most equilibrated result, in the sense of equivalent Recall and Precision was obtained with: Kp=1.8(or 1.9); Mean_Factor = 3.0 and β =0.8
which give us a recall of 72.2 and a precision of 72.2.
The best precision (79 for a recall of 64) was obtained with:
Kp=3.0; Mean_Factor = 5.0 and β =0.7
and the best recall (75 for a precision of 61) was obtained with:
Kp=1.0; Mean_Factor = 2.0 and β =0.9

Our TRECVid 2004 submission consists in the same setups than explained before:
The most equilibrated result is a recall of **70.2** and a precision of **63.4**.
The best precision is **73** for a recall of **65** and the best recall is **74** for a precision of **57**.

However, the rough indexing framework without classification of transitions effects, we call it non-classified (NC) bound, shows that performances are 0.83 of Recall and Precision on TRECVID2003, and 0.82 of Recall and 0.79 of Precision on TRECVID2004.

The improvement of our method to attain the NC bound are possible and being implemented now. New fusion models of low-level indexes, which are developed now, can also improve the NC bound. The performances on TREC 2004 are 3 times faster than real time, which is the 4th best complexity among the 19 participants. The three algorithms of best complexity have also best Recall/Precision than our algorithm.

The following graph shows our Recall/Precision position compared to the other attendees.



Recall and Precisions for All Transitions

6. Bibliographie :

[1] D. A. Adjeroh , M. C. Lee , C. U. Orji, *A principled approach to fast partitioning of uncompressed video*, Proceedings of the 1996 International Workshop on Multi-Media Database Management Systems (IW-MMDBMS '96), p.115, August 14-16, 1996

[2] J.S. Boreczky and L.A. Rowe, *A Comparison of Video Shot Boundary Detection Techniques,* Journal of Electronic Imaging, 5(2), April 1996.

[3] HongJiang Zhang , Atreyi Kankanhalli , Stephen W. Smoliar, *Automatic partitioning of full-motion video*, Multimedia Systems, v.1 n.1, p.10-28, June 1993

[4] A. Nagasaka and T. Tanaka. *Automatic video indexing and full-video search for object appearances*. In IFIP Proc. of Visual Database Systems, pages 113--127, 1992.

[5] Rainer Lienhart. *Reliable Transition Detection In Videos: A Survey and Practitioner's Guide*" International Journal of Image and Graphics (IJIG), Vol. 1, No. 3, pp. 469-486, 2001.

[6] G. M. Quenot, D. Moraru, and L. Besacier, CLIPS at TRECVID : *Shot Boundary Detection and Feature Detection*, TRECVid 2003, NIST, Gaithersburg, MD, USA, 17-18 November 2003.

[7] C.-Y. Lin, M. Naphade, A.P. Natsev, B. Tseng, Y. Wu, D. Zhang, G. Iyengar, C. Neti, H. Nock, A. Amir, M. Berg, W. Hsu, *IBM Research TRECVID-2003 Video Retrieval System,* TRECVid 2003, NIST, Gaithersburg, MD, USA, 17-18 November 2003.

[8] Xian-Sheng HUA, Pei YIN, Huajian WANG, Junfeng CHEN, Lie LU, Mingjing LI, Hong-Jiang ZHANG. *MSR-Asia at TREC-11 Video Track*. TREC Video Retrieval Evaluation (TRECVID 2002). 2002.

[9] Farshid Arman , Arding Hsu , Ming-Yee Chiu, *Image processing on compressed data for large video databases*, Proceedings of the first ACM international conference on Multimedia, p.267-272, August 02-06, 1993, Anaheim, California, United States

[10] X. Huang, G. Wei, and V. A. Petrushin, Accenture Technology Laboratories (ATL), *Shot Boundary Detection and High-Level Features Extraction for the TREC Video Evaluation 2003*

[11] M. Sugano, K. Hoashi, K. Matsumoto, and Y. Nakajima, KDDI R&D Laboratories Inc., *Shot Boundary Determination on MPEC Compressed Domain and Story Segmentation Experiments for TRECVID 2003*.

[12] Alan Hanjalic, Member, IEEE, *Shot-Boundary Detection: Unraveled and Resolved?*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 12, NO. 2, FEBRUARY 2002

[13]. F. Manerba, J. Benois-Pineau, R. Leonardi, *Extraction of foreground objects from a MPEG2 video stream in rough indexing framework*, In Proc. Storage and Retrieval Methods and Applications for Multimedia 2004, EI'2004 SPIE, San Jose (CA) 18-22 January, 2004.

[14] M. Durik and J. Benois-Pineau, *Robust motion characterisation for video indexing based on MPEG2 optical flow*, Proc. CBMI'2001, 2001, 57-64.

[15] P. Krämer, J. Benois-Pineau, and J.-P. Domenger, *Scene Similarity Measure for Video Content Segmentation in the Framework of Rough Indexing Paradigm*. 2nd International Workshop on Adaptive Multimedia Retrieval AMR 2004, August 2004, Valencia, Spain