IIT / NCSR “Demokritos” at TRECVID 2006: **SHOT BOUNDARY DETECTION**

I. Pratikakis, I. Dounis, B.Gatos and S. Perantonis

Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
P.O. BOX 60228, GR-153 10 Agia Paraskevi, Athens, Greece.
{ipratika, jdounis, bgat, sper} @iit.demokritos.gr

Abstract

In this notebook paper, we summarize our results in the framework of TRECVID 2006 competition. In particular, we have produced results for the Shot Boundary detection task for which we will describe the submitted runs and analyze their performance.
The applied methodology relies on spatial segmentation and a similarity measure based on Earth Mover’s Distance that produces signatures for a given spatio-temporal support. We have used alternative signatures for abrupt cuts and gradual transitions.

1. Introduction

An important step towards search and retrieval in videos is the boundary detection in shots. In a video stream, the scene transition between two shots can be of two main types, abrupt or gradual. *Abrupt scene changes* appear, when in a series of frames the frame f_i belongs to one scene and the subsequent frame f_{i+1} belongs to the next scene. Abrupt scene change is also called a cut, similar to the way a scene is cut in order for the next scene to appear. *Gradual shot changes*, on the other hand, can be composted of more complicated effects for the transition. According to the way one scene disappears and the other scene appears, gradual shot transitions can be one of the following types: wipe, fade and mix/ dissolve/ crossfade. Although the types of gradual shot transitions are many, one global rule exists and involves several succeeding frames. Frame f_i belongs to one scene, frame f_{i+N} belongs to the next scene and the N - 1 frames in between represent the transition between the two scenes.

The basic components in our methodology are (i) spatial segmentation [Deng-2001]; (ii) a similarity measure between frames based on Earth Mover’s Distance [Rubner-2000] and (iii) separate transition modeling for both abrupt cuts and gradual transitions. In particular, we apply a spatial segmentation at each frame of the sequence, thus producing a set of regions at this frame. For each region, we compute a feature set which is going to feed the required feature distribution for each region in the Earth Mover’s distance. In TRECVID 2006 experiments, we have used 2 different feature sets applied on 4 runs each of them. Specifically, we have used the FS1 (Mean RGB color, Mean adjacent RGB color, Center of mass, Mean adjacent gradient) and the FS2 (Mean RGB color, Center of mass, Mean adjacent gradient) (see Table 1 - Parameters).

After the computation of the Earth Mover's Distance between consecutive frame pairs we get a 1-d similarity histogram which is the base for the next step towards shot boundary detection for both cuts types. This will be described in the following Sections which are dedicated to the detection of the particular cuts type, namely, abrupt cut and gradual cut.

2. Abrupt Cut Detection

The computed 1-d similarity histogram is a first estimation of the adjacent frame similarity in a video sequence. In the ideal case, the boundaries in abrupt cuts would be distinct spikes in the histogram. Thus, applying a threshold would be easy to get the desired boundaries. Nevertheless, in a real situation, either noise or lack of the selected feature set to grasp the semantic difference between frames has led us to go for certain steps that will refine the result from the 1-d similarity histogram. These steps are described in the following :

Step 1 : We apply the principle of the *dynamics of maxima* at the 1-d landscape. This will enable us to provide a meaningful hierarchy at the local maxima of the 1-d histogram, thus, applying a threshold we will get candidate shot boundaries which are detected in terms of the global information of the sequence.

Step 2 : When we apply, a threshold *Tc* (see Table 1 - Parameters) there are false positives produced by flashes, etc. that we need to reject. For this, we introduce a modeling of an abrupt cut in terms of the EMD. For this modeling, we compute for any candidate boundary a set of similarities based on EMD between the current frame and each of the S previous frames. In our case, S=5. For a particular frame *n*, this is applied on the next five (5) frames, till *+5*. According to this, we get a spatiotemporal template which expresses a linear dissimilarity decrease in time. An example of such a template is shown in Figure 1. Finally, among the candidate shot boundaries, we keep only those that fit the proposed template.

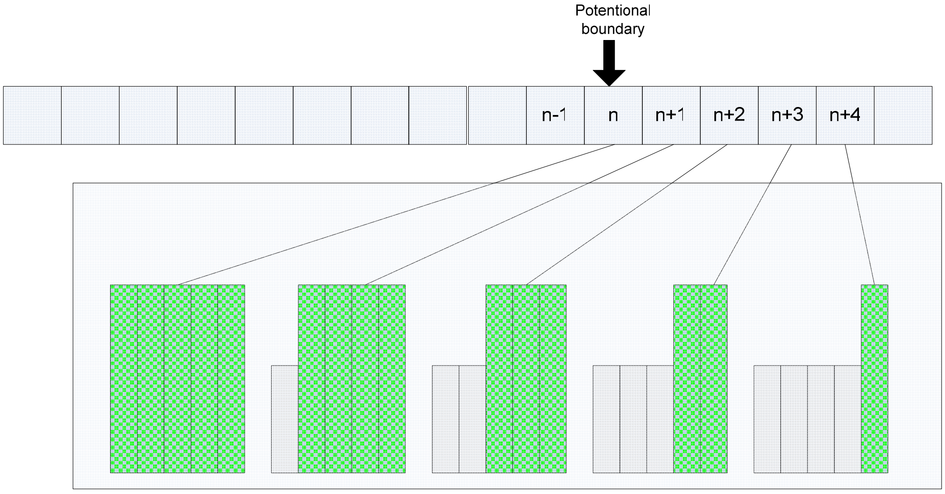

Figure 1: Spatio-temporal template for Abrupt cuts detection

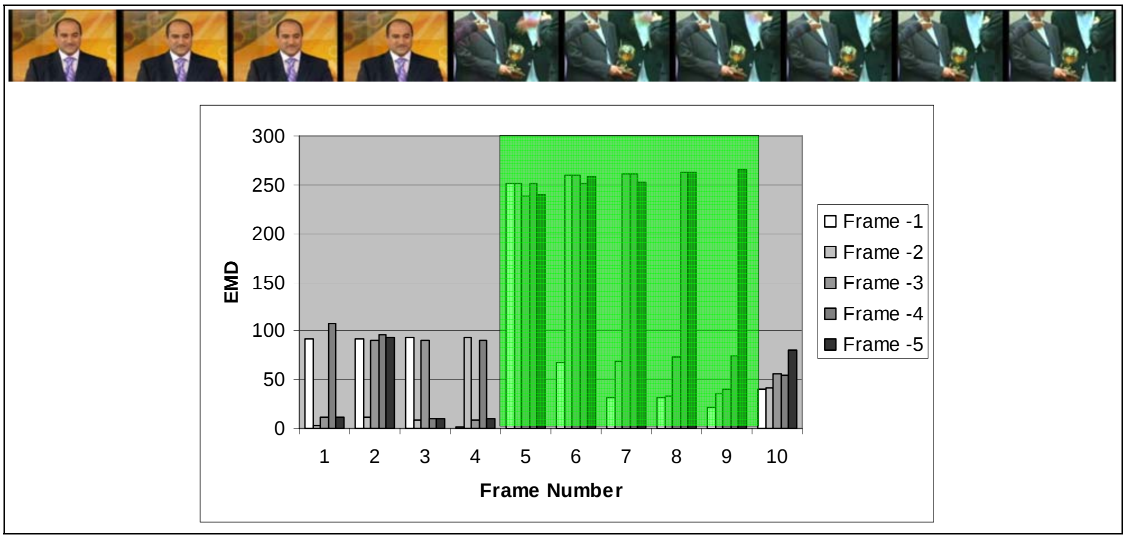

Figure 2: Example of detected abrupt cut. Frames that fit the spatio-temporal template for abrupt cuts are marked in the graph.

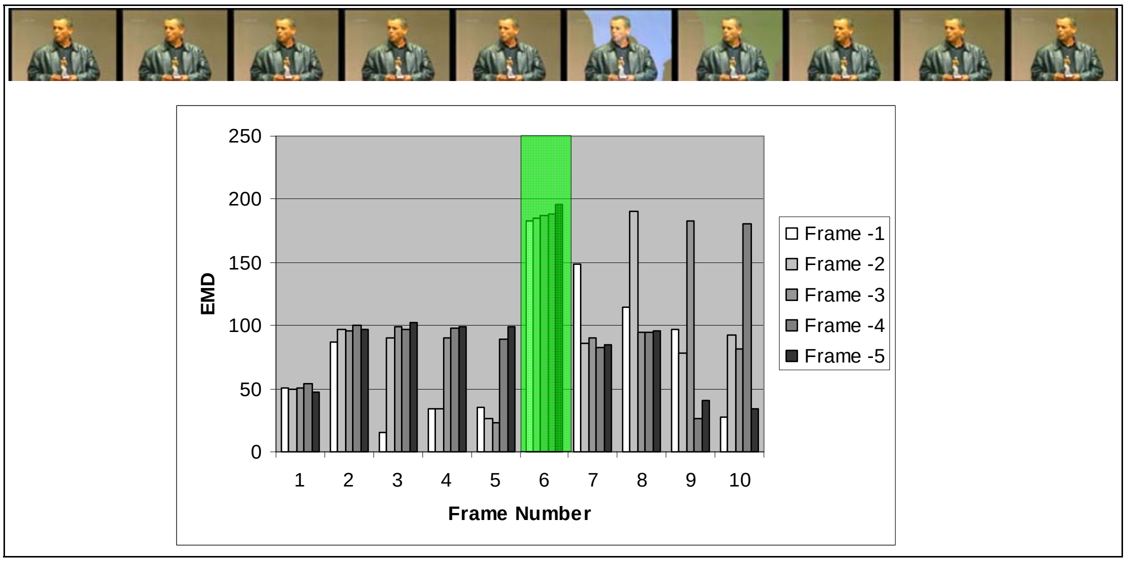

Figure 3 : Example of filtering a flash. In this case, frames 6-10 do not fit the spatio-temporal template for abrupt cuts and the candidate boundary is discarded.

3. Gradual transition detection

In the case of gradual transition detection, we follow three (3) consecutive steps till we reach the final outcome. Here also, the prominent underlying principle is the fit to a spatio-temporal template of each frame in the video sequence. These steps are described in the following :

Step 1 :

First of all, we consider that all frames have been segmented and we have calculated the aforementioned feature sets (FS1 or FS2) for each region in every frame. Then, the goal is to decide whether a frame belongs to gradual transition area or not. We opt for a binary decision. This decision is based on the fit of a frame *n* to a spatiotemporal template that corresponds to a histogram by taking consecutively the EMD between frame *n* and each of *n+i* frames, with *i* in the interval [1,..,S]. In our experiments S=5. Thus, a histogram with 5 bins is constructed for which, in the ideal situation, it is expected that it follows a monotonously increasing valuation.
The fit gives us a candidate frame to belong to a gradual transition.

Step 2:

A merging process is applied to compensate for outliers of the proposed template. Merging occurs only if there exist areas between candidate gradual transition areas with not more than 3 frames.

Step 3 :

After the merging process we expand the left (lower frame) boundary of the candidate gradual transition area by S frames to compensate the fact that the pattern begins S frames after the lower transition boundary.

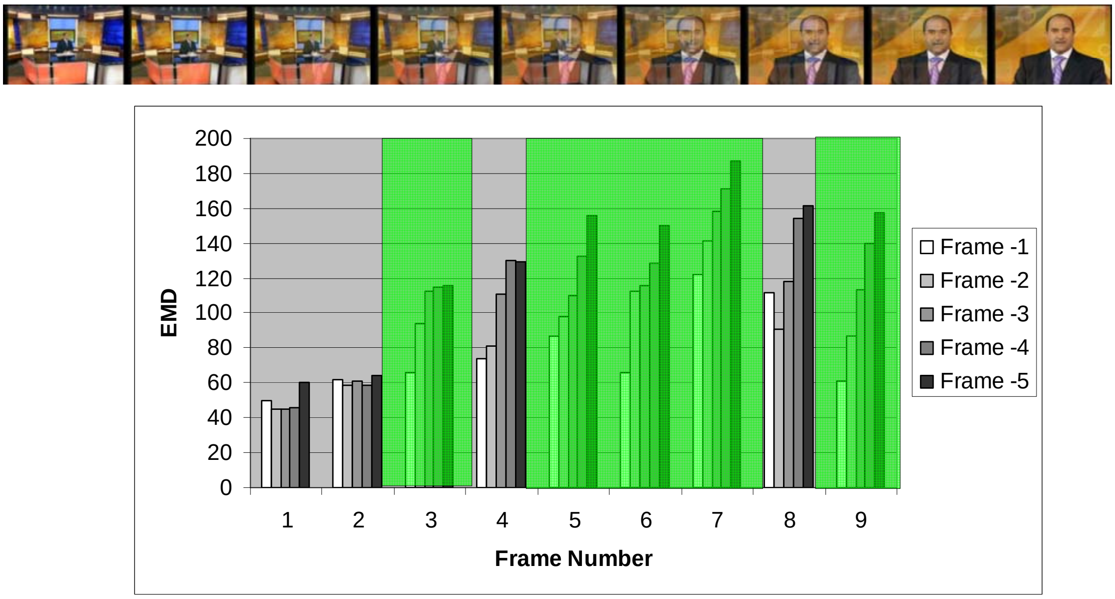

Figure 4 Example of gradual transition (FOI) detection. Frames that fit the spatio-temporal template for gradual transitions are marked in the graph.

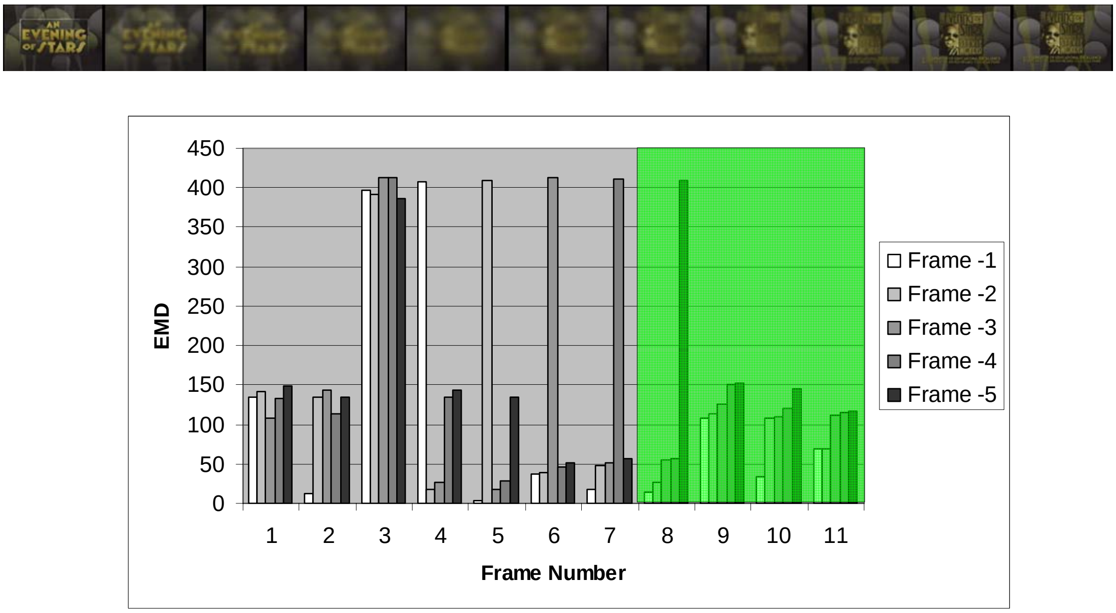

Figure 5: Example of gradual transition detection (Dissolve). Frames that fit the spatio-temporal template for gradual transitions are marked in the graph. In this example, we exemplify the need of using Step 3 as it is explained at Section 3.

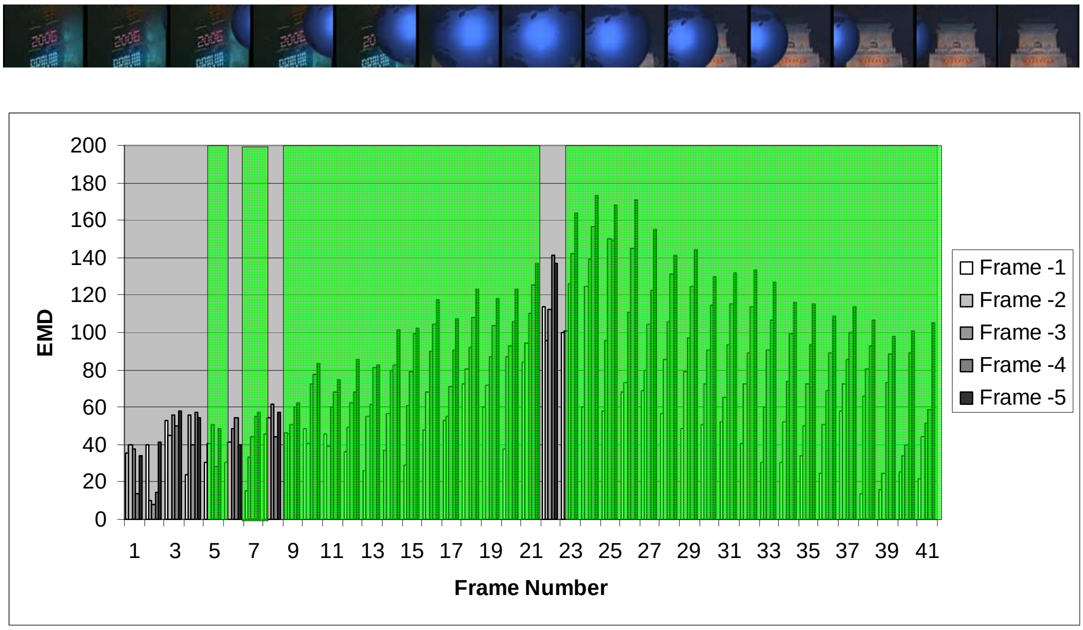

Figure 6: Example of gradual transition detection (Wipe)**.** Frames that fit the spatio-temporal template for gradual transitions are marked in the graph.

4. Experimental results

At Table 1, it is shown the detailed performance of our proposed methodology.
It is obvious that although abrupt cuts detection is satisfactory, the gradual transition detection is very weak. We strongly believe that incorporating explicit motion information we will dramatically improve our results.

Run #	Parameters	All		Abrupt Cuts		Gradual Transitions			
		Recall	Precision	Recall	Precision	Recall	Precision	Frame Recall	Frame Precision
1	FS1 Tc = 80 Tg = 100	0.585	0.684	0.742	0.694	0.161	0.587	0.584	0.859
2	FS1 Tc = 60 Tg = 100	0.637	0.561	0.819	0.559	0.148	0.590	0.577	0.854
3	FS1 Tc = 80 Tg = 120	0.578	0.689	0.755	0.689	0.100	0.689	0.662	0.884
4	FS1 Tc = 60 Tg = 120	0.624	0.575	0.822	0.571	0.092	0.701	0.664	0.878
5	FS2 Tc = 80 Tg = 100	0.556	0.695	0.716	0.710	0.126	0.524	0.480	0.796
6	FS2 Tc = 60 Tg = 100	0.620	0.587	0.810	0.591	0.109	0.516	0.482	0.772
7	FS2 Tc = 80 Tg = 120	0.553	0.699	0.727	0.706	0.086	0.583	0.484	0.778
8	FS2 Tc = 60 Tg = 120	0.615	0.586	0.815	0.589	0.075	0.520	0.481	0.750

Table 1: Detailed quantitative analysis of our experiments

References

[Rubner-2000] Y., Rubner and C., Tomasi. *Perceptual metrics for image database navigation*. Kluwer Academic Publishers, Boston, 2000.
[Deng-2001] Y. Deng and B.S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001