University of Modena and Reggio Emilia at TRECVID 2006

Costantino Grana, Roberto Vezzani, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
{grana.costantino, vezzani.roberto, cucchiara.rita}@unimore.it

Structured Abstract

What approach or combination of approaches did you test in each of your submitted runs?
TRECVID2005_UNIMORE_??.xml: the same linear transition detector (LTD) was tested for every run, with ten uniformly spaced thresholds for the detection.
What if any significant differences (in terms of what measures) did you find among the runs?
The system behaved as expected: the higher the threshold the better the recall. Of course the precision lowered correspondently. Interesting enough, it seems that we cannot overcome the overall limit around 80% for recall and 88% for precision, independently of the other parameter.
Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?
One of the main objective of our system was to test the performance of a single algorithm for both cuts and gradual transitions. So all the merit and the demerits are related to our LTD.
Overall, what did you learn about runs/approaches and the research question(s) that motivated them?
The use of a single algorithm allows the system to be run without training. Just a single parameter may be employed to tune the sensibility of the system, thus allowing its use in general purpose/user friendly systems.

1. Introduction

This is the second year that the University of Modena and Reggio Emilia tries the Shot Boundary Detection task of TRECVID. As in last year try [1], our approach is strictly focused on gradual transitions with a linear behavior, including cuts. We developed an iterative algorithm that, given a range of frames possibly including a transition, alternatively tries to find the best center position or the best length, by minimizing an error function, which measures the fitness of data to the linear model [2]. The features used to characterize the frames are the DC color image in the RGB color space and the three RGB independent histograms computed on the full sized image.

2. Shot Boundary Detection

Before describing our algorithm in detail, it is useful to define the ideal model of linear transition and to underline its important properties. These will be exploited by the algorithm to cope with non idealities and to measure the confidence of the detection.

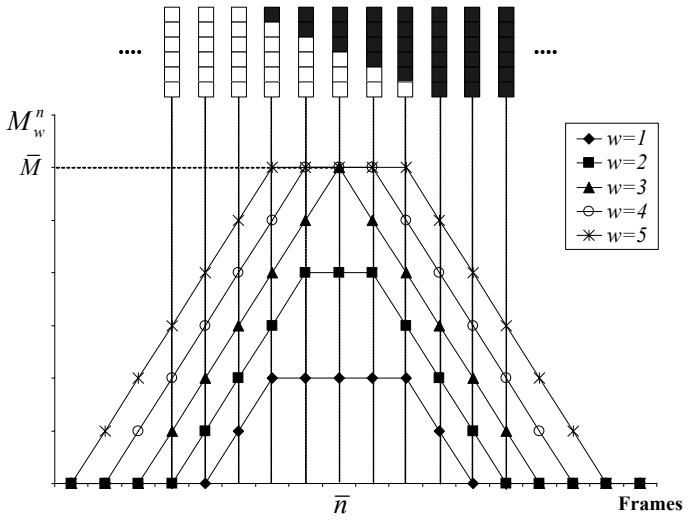

Figure 1. Values of M_w^n for an ideal linear transition with $L=5$ at varying *w*.

2.1. The Transition Model

Let's consider two consecutive shots in a video sequence, the first one ending at frame *e*, and the second one starting at frame *s*, with $e<s$. If $s=e+1$ we have an abrupt cut, otherwise there are some frames of gradual transitions between *e* and *s*.

To design a shot segmentation algorithm, two assumptions must be done: the first one is that a feature $F(t)$ is computable for each frame at time *t*, with the characteristic of being discriminating and almost constant within the shot; ideally

$$\begin{aligned} &F(t)=F(e),\forall t\le e \\ &F(t)=F(s),\forall t\ge s \\ &F(e)\neq F(s) \end{aligned} \tag{1}$$

The second assumption is that a distance function exists in the feature space Φ: $d:\Phi\times\Phi\to\mathbb{R}$, which shows a constant behavior during the transition. Ideally:

$$d\left(F(t),F(t-1)\right)=c \quad e<t\le s \tag{2}$$

Sometimes there is confusion on the definition of length of a transition, because one may include in the count the first frame of the new shot after the transition (e.g. [3]), or the last one of the previous one. In our model, the length is the number of frames in which the transition is visible, that is $L=s-e-1$. Note that this model includes in the definition of transition abrupt cuts too, as transitions with length $L=0$. The transition center is defined as $\bar{n}=(e+s)/2$ and may correspond to a non-integer value, that is an inter-frame position. This is always an inter-frame position in case of cuts.

Differently from other difference metric formulations, instead of computing the difference between the frames $F(i)$ and $F(i+w)$, with *w* being the *frame-step*, we calculate a metric M_w^n centered on frame or half-frame *n*, with $2n\in\mathbb{N}$, and with frame-step $2w\in\mathbb{N}$. It is defined as:

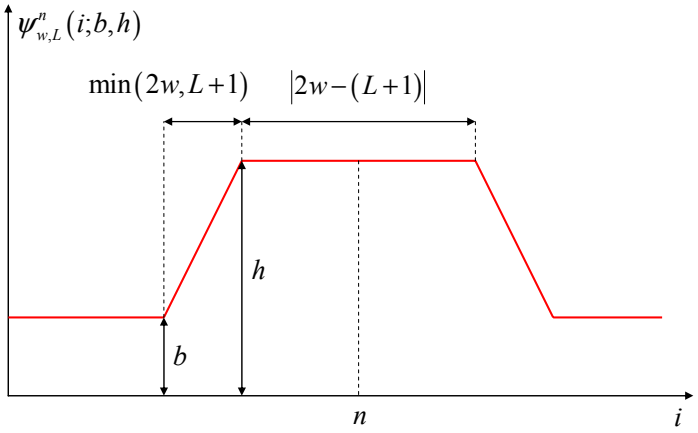

Figure 2. Trapezoidal shaped function $\psi^n_{w,L}(i;b,h)$

$$M^n_w = \begin{cases} d\left[F(n-w), F(n+w)\right] & n+w \in \mathbb{N} \\ \frac{1}{2}\left[M^{n-\frac{1}{2}}_w + M^{n+\frac{1}{2}}_w\right] & otherwise \end{cases} \tag{3}$$

The second term of the expression is a linear interpolation adopted for inter-frame positions. This is necessary because the feature F is relative to a single frame and cannot be computed at half-frames.

In Fig. 1 we see an example of an ideal linear transition with $L=5$, from a shot with white pixels to one with black pixels. If the transition is perfectly linear according with the hypothesis of Eq. 1 and Eq. 2, the shape of function M^n_w is an isosceles trapezoid centered in $\bar{n}$, for each w, that degenerates into a triangle when $2w = L+1$.

We can verify that in this ideal case, given the model and Eq. 3, both the up and down slopes last for $\min(2w, L+1)$ frames, and that the plateau of absolute maximum is $|2w-(L+1)|$ long. It's also straightforward to verify that:

$$M^{\bar{n}}_w < \bar{M}, \text{if } 2w < L+1; \qquad M^{\bar{n}}_w = \bar{M}, \text{if } 2w \geq L+1 \tag{4}$$

where $\bar{M} = \max_{w,n} M^n_w$ (see Fig. 1). We define $\psi^n_{w,L}(i;b,h)$ the generic trapezoidal function, centered in n, whose value is h at the center (the absolute height of the minor base) and b is the value outside the trapezoid. The function is plotted in Fig. 2. We define $\psi^n_{w,L}(i) = \psi^n_{w,L}(i;0,M^n_w)$, the function which corresponds to the ideal transition case.

In the real case, camera and objects motion, color and luminance variation and so on cause the feature F to be non constant on the shot, thus making Eq. 1 and Eq. 2 not satisfied. The consequence is that the shapes of both the slopes and the plateau are usually disturbed.

2.2. Two-steps Algorithm

Due to lack of ideality in most of the shot transitions, instead of relying only on correlation between data and the ideal $\psi^n_{w,L}(i)$ function, we employ an algorithm constructed of two steps: the first one searches for the transition center position n, assuming a fixed frame step $2w$, and the second searches for the transition length L, by trying different values of w, but keeping the

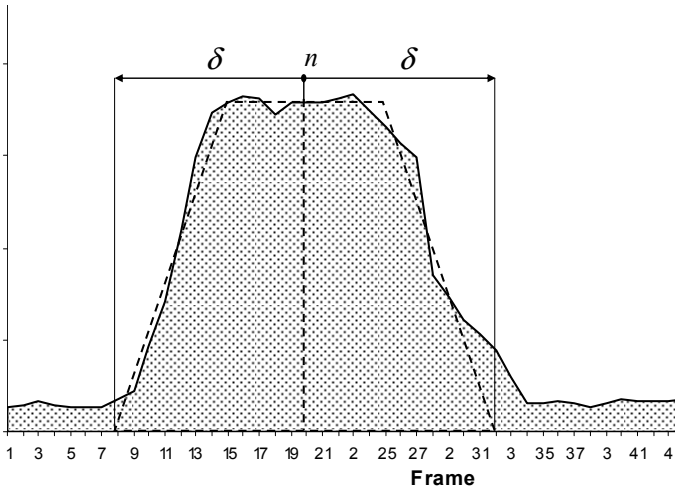

Figure 3. Example of real M_w^n values and the best trapezoid fitted.

transition center fixed. While in the ideal case even the first step would be sufficient, in real cases an error in locating the center position would also lead to a wrong estimate of the length. For this reason a second step is introduced to provide a different view of the function behavior, a possible confirmation on the first step outcome and a new estimate for the window size. Iteratively repeating the two steps allows progressively decreasing the error. In this section we explain in details our transition detection algorithm. We perform the following analysis on overlapped windows of 60 frames, distant 30 frames each other, since we suppose that transitions are much shorter and farther than that.

2.2.1. First step.

In the first step the values of M_w^n are calculated using the frame-step $\bar{w}$, which is found in the previous iteration of the algorithm, or it's arbitrary chosen for the first iteration. The best trapezoid $\psi_{w,L}^n(i)$ is searched by moving the center *n*, and trying different values for *L,* but keeping $\bar{w}$ fixed. The trapezoid extends over $\delta = \min(2w, L+1) + |w - (L+1)/2|$ frames on the left and on the right of the center frame. For each couple of *n* and *L* the following matching measure is computed:

$$\Lambda_{\bar{w},L}^n = \sum_{i=n-\delta}^{n+\delta} \min\left(M_{\bar{w}}^i, \psi_{\bar{w},L}^n(i)\right) - \sum_{i=n-\delta}^{n+\delta} \left| M_{\bar{w}}^i - \psi_{\bar{w},L}^n(i) \right| \tag{5}$$

The value of *n* is searched within the 60 frames window, and also *L* must be selected such that $n+\delta$ and $n-\delta$ don't exceed the window.

In Eq. 5, two components are evident: the first one is needed to maximize the area under the trapezoid, while the second component describes the similarity of our linear hypothesis with the data. It is very important to include both components, since we expect the distance measure to give a trapezoidal shape (the second term in Eq. 5), but we also request its *strength*, i.e. the amount of difference between the first and the second scene, to be significant. The first term in Eq. 5 in fact describes how much the value of M_w^n surpasses the ideal trapezoid. After finding

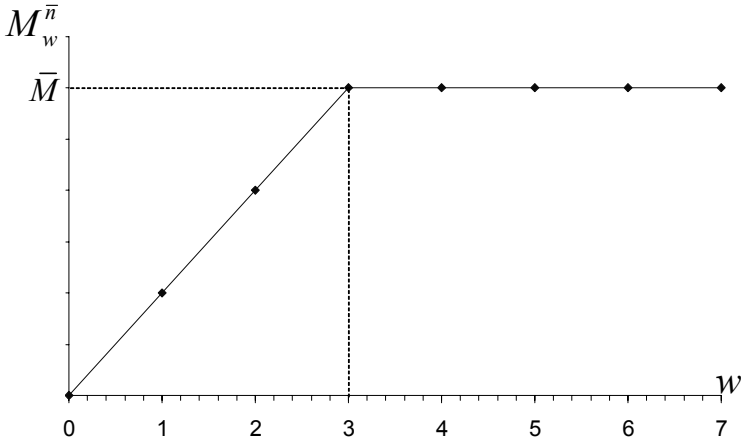

Fig. 4. Values of the distance metric M_w^n, with respect to different *w* values. This corresponds to the transition of Fig. 1.

the trapezoid which maximizes $\Lambda_{\bar{w},L}^n$, we consider $\bar{n} = \arg\max \Lambda_{\bar{w},L}^n$ the candidate transition center. In Fig. 3 we show an example of trapezoid fitting with real data.

2.2.2. Second Step.

Thanks to the definition of M_w^n as a distance function centered in *n*, as in Eq. 3, increasing the frame-step *w* makes the value of M_w^n to grow up to an absolute maximum when $w = (L+1)/2$ and then to be stable. It is easy to demonstrate that, in the ideal case, this growth is linear. Thus the growing function can be plotted as shown in Fig. 4, with a linear slope followed by a horizontal line, when the value of M_w^n is stable. The second step of the algorithm uses this propriety to give an estimate of the transition length, by finding the smallest *w* which maximizes M_w^n. To provide a technique able to deal with noise, the tilt change of the chart is searched by minimizing the function:

$$Z_w^{\bar{n}} = \sum_{i=0}^{w} \left| M_i^{\bar{n}} - \frac{M_w^{\bar{n}}}{w} i \right| + \sum_{i=w+1}^{W} \left| M_i^{\bar{n}} - M_w^{\bar{n}} \right| \tag{6}$$

where *W* is the maximum size that a transition can assume. The *w* value that minimizes $Z_w^{\bar{n}}$ becomes our current frame step for the next iteration of the algorithm.

In simple cases the algorithm progressively narrows the trapezoid minor base leading to the expected triangular shape. Convergence is not guaranteed in non ideal conditions, and, for this reason, we add a convergence constraint: at each iteration the minor base of $\psi_{w,L}^n(i)$ is forced to become smaller. In Fig. 5 the M_w^n values are shown for 4 successive iterations of the algorithm in a real gradual transition case. At each iteration, we achieve a more precise estimate of the

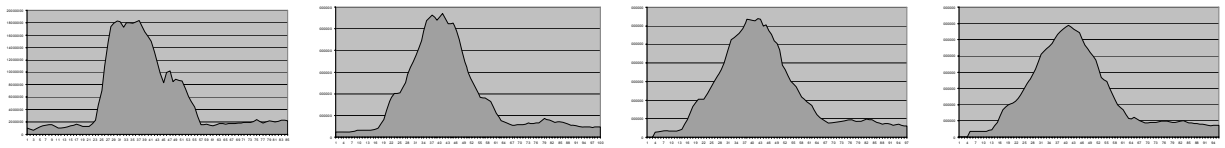

Figure 5. Four successive iterations of the algorithm in a real gradual transition: at each iteration, the shape of M_w^n values becomes more similar to a triangle

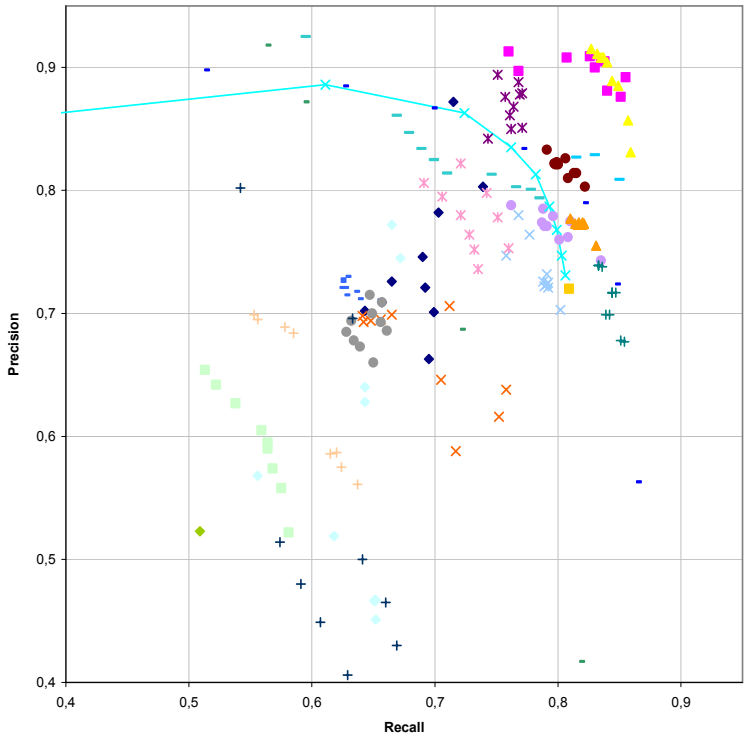

Fig. 6. Graph reporting the results obtained by the TRECVID participants. Some results are out of scope. Our results are shown with a cross connected by a line.

transition center and length, and thus a shape more similar to a triangle.

2.2.3. Decision Space.

Given the transition length $L = 2w - 1$ and its center $\bar{n}$, as detected by the algorithm, the function $\psi^{n}_{w,L}(i)$ becomes triangular shaped. We must now verify the significance of the transition and how much the real data fit to the linear transition model. We introduce the following measure:

$$Peak^{\bar{n}}_{\bar{w}} = M^{\bar{n}}_{\bar{w}} - \min\left(M^{\bar{n}-2\bar{w}}_{\bar{w}}, M^{\bar{n}+2\bar{w}}_{\bar{w}}\right). \tag{7}$$

The Peak value measures the height of the center value with respect to the lower of the two values of M in correspondence to the extremes of the triangle, and provides information on the transition significance. In fact, while in the model $M^{n\pm 2w}_{w} = 0$, in real cases this is not true, because of object and camera motion that causes the feature F to be not constant before and after the transition. To cope with this we have to get rid of the hypothesis of having an isosceles triangle and define the fitting error measure as:

$$err^{\bar{n}}_{\bar{w}} = \frac{1}{4\bar{w}} \sum_{i=1}^{2\bar{w}} \left| M^{\bar{n}-i}_{\bar{w}} - \psi^{\bar{n}}_{\bar{w},L}(\bar{n}-i, M^{\bar{n}-2\bar{w}}_{\bar{w}}, M^{\bar{n}}_{\bar{w}}) \right| + \left| M^{\bar{n}+i}_{\bar{w}} - \psi^{\bar{n}}_{\bar{w},L}(\bar{n}+i, M^{\bar{n}+2\bar{w}}_{\bar{w}}, M^{\bar{n}}_{\bar{w}}) \right| \tag{8}$$

The error sum is divided by the triangle's base $4\bar{w}$ to obtain a measure which is independent from the transition length. The final decision space is then based on two parameters only, which are the same for cuts and transitions. The decision rule to have a valid transition is:

$$\alpha \cdot err_{\bar{w}}^{\bar{n}} + \beta \cdot Peak_{\bar{w}}^{\bar{n}} > 1 \tag{9}$$

where α and β are two real coefficients.

3. Results

In Fig. 6 the results obtained by our system with respect to all other TRECVID participants are shown. For all the ten submitted runs, the error coefficient α was fixed to -32, while the peak coefficient β was varied from 0.0001 to 0.001 in ten equally spaced steps. As expected, the lower the coefficient, the better the recall. Of course the precision lowered correspondently. Interesting enough, it seems that we cannot overcome the overall limit around 81% for recall and 88% for precision, independently of the other parameters. Another noticeable result is given by run TRECVID2005_UNIMORE_01.xml (peak value 0.0001), in which the high request for the peak value does not lead to a better precision. In fact the detected transitions with higher peak are, instead false ones.

Reference

[1] Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. Ali, M. Shah, C. Grana, and R. Cucchiara, "Video understanding and content-based retrieval," in *TREC Video Retrieval Evaluation WorkshopOnline Proceedings (TRECVID2005)*, Gaithersburg, MD, USA, Nov. 2005. [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/ucf.pdf

[2] C. Grana, G. Tardini, R. Cucchiara, "MPEG-7 Compliant Shot Detection in Sport Videos" in *Proceedings of IEEE International Symposium on Multimedia (ISM2005)*, Irvine, California, USA, pp. 395-402, December 12-14, 2005.

[3] Bescos, J., Cisneros, G., Martinez, J.M., Menendez, J.M., Cabrera, J. A unified model for techniques on video-shot transition detection. *IEEE Transactions on Multimedia*, 7, 2 (Apr. 2005), 293-307.