An Unified Transition Detection Based on Bipartite Graph Matching Approach

Zenilton Kleber Gonçalves do Patrocínio Jr., Silvio Jamil F. Guimarães,
Henrique Batista da Silva, and Kleber Jacques Ferreira de Souza

Audio-Visual Information Processing Laboratory (VIPLAB)
Institute of Informatics - Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brazil
{zenilton,sjamil,henriquebat,kleberjac}@pucminas.br
http://www.viplab.inf.pucminas.br

Abstract. This paper addresses transition detection which consists in identifying the boundary between consecutive shots. In this work, we propose an approach to cope with transition detection in which we define and use a new dissimilarity measure based on the size of the maximum cardinality matching calculated using a bipartite graph with respect to a sliding window. The experiments have used two video datasets which presents a variety of different video genres with 3079 transitions. Our method achieves performance measures similar to the best results found in the literature with a much simpler classification approach.

Keywords: Bipartite graph matching, cut, gradual transition.

1 Introduction

An hierarchical model for video analysis and segmentation is usually divided into four levels based on its temporal resolution. At the lowest level one can find the most basic unit, i.e., a single video frame. Several of those frames are gathered into a shot that represents a continuous camera recording. Some shots present a storytelling coherence and they are grouped into distinct scenes. Finally an assembly of different scenes constitute a digital video. Amongst the problems related to video analysis and indexing, sometimes video segmentation can be considered as an essential first step. This paper addresses transition detection which is part of video segmentation problem, and consists in identifying the boundary between consecutive shots. The most common approach to cope with transition detection is based on the use of a dissimilarity measure [1]. A review of the most popular methods for cut (abrupt transition) detection (such as pixel-wise comparison, histogram comparison, etc) can be found in [2, 3]. If two frames belong to the same shot, then their dissimilarity measure should be small. Two frames belonging to different shots generally yield a high dissimilarity measure. In the same way, a dissimilarity measure concerning the frames of a gradual transition is difficult to define and the quality of this measure is very important for the whole segmentation process.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 184–192, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Another approach to the video segmentation problem is to transform the video into a 2D image [4], and apply image processing methods on this image to extract the different patterns related to each transition. Some works on gradual transitions detection can be found in [5–7]. Zabih et al. [5] proposed a method based on edge detection which is very costly due to the computation of edges for each frame of the sequence. Fernando et al. [6] used a statistical approach that considers features of the luminance signal. This approach presents high precision on long fades. Zhang et al. [7] introduced the twin-comparison method in which two different thresholds are considered. In [4], Ngo et al. applied Markov models for shot transition detection which fails in the presence of low contrast between textures of consecutive shots. Recently, Bescós et al. [8] proposed a unified framework with very good results for detecting both cuts and gradual transitions. The major drawback of this method is the large number of parameters (thresholding) that are needed to adjust the classification algorithm. Finally, Grana et al. [9] proposes a linear transition detector for both cuts and linear gradual transitions. Their method searches for the transition center and transition length using different values of frame step. However, this algorithm assumed that the feature information is computable, discriminating, and constant within the shots. In this work, we propose an unified approach to cope with transition detection. In this paper, we define and use a new dissimilarity measure based on the size of the maximum cardinality matching calculated using a bipartite graph with respect to a sliding window. In [10], an approach based on a bipartite graph was used only for cut detection in which a dissimilarity measure between two consecutive frames was calculated from maximum cardinality matching. The main contribution of this work is the application of a new simple and efficient dissimilarity measure unifiedly to solve the cut and gradual transition detection problem.

This paper is organized as follows. In Section 2 we define a new dissimilarity measure and the transition detection problem resolution using that measure. In Section 3 our methodology is fully presented. In Section 4 we perform an analysis for transition detection involving our method using three different quality measures. Some conclusions and a summary of future works are given in Section 5.

2 A Dissimilarity Measure

In [10], it was defined some concepts used here, like point similarity graph and list of frame points. Unformally, the point similarity graph $\mathrm{G}^{\delta,\lambda}(t_1, t_2)$ is created from a list of frame points, L_{t_1} and L_{t_2}, computed from a visual rhythm which is a simplification of the frame. The graph vertex is the frame points (pixels) and the weighted edge is the similarity value between two points.

Definition 1 (Matching – $\mathrm{M}^{\delta,\lambda}$). *Let* $\mathrm{G}^{\delta,\lambda}(t_1, t_2)$ *be a point similarity graph between the frames* f_{t_1} *and* f_{t_2} *represented by their list of frame points* L_{t_1} *and* L_{t_2} *[10]. A subset* $\mathrm{M}^{\delta,\lambda} \subseteq \mathrm{E}^{\delta,\lambda}$ *is a matching if any two edges in* $\mathrm{M}^{\delta,\lambda}$ *are not adjacent. The size of matching* $\mathrm{M}^{\delta,\lambda}$*,* $|\mathrm{M}^{\delta,\lambda}|$*, is the number of edges in* $\mathrm{M}^{\delta,\lambda}$*.*

Definition 2 (Maximum cardinality matching – $\overline{\mathrm{M}^{\delta,\lambda}}$). *Let* $\overline{\mathrm{M}^{\delta,\lambda}}$ *be a matching in a point similarity graph* $\mathrm{G}^{\delta,\lambda}(t_1, t_2)$*. So,* $\overline{\mathrm{M}^{\delta,\lambda}}$ *is the maximum* *cardinality matching if there is no other matching* $\mathrm{M}^{\delta,\lambda}$ *in* $\mathrm{G}^{\delta,\lambda}(t_1, t_2)$ *such that* $|\mathrm{M}^{\delta,\lambda}| > |\overline{\mathrm{M}^{\delta,\lambda}}|$. Solving the maximum cardinality matching on a bipartite graph could done with $O(E\sqrt{V})$ operations, in which V and E represent the number of nodes and edges, respectively. Based on the size of maximum cardinality matching we can define an interframe dissimilarity measure in the following manner:

Definition 3 (Dissimilarity measure – $\mathrm{DIS}^{\delta,\lambda}(t_1, t_2)$**).** *Let* $\overline{\mathrm{M}^{\delta,\lambda}}$ *be a maximum cardinality matching in a point similarity graph* $\mathrm{G}^{\delta,\lambda}(t_1, t_2)$*. So, the dissimilarity measure* $\mathrm{DIS}^{\delta,\lambda}(t_1, t_2)$ *can be calculated as* $\mathrm{DIS}^{\delta,\lambda}(t_1, t_2) = 1 - \frac{|\overline{\mathrm{M}^{\delta,\lambda}}|}{\max\{|\mathrm{L}_{t_1}|, |\mathrm{L}_{t_2}|\}}$.

Two consecutive frames that are similar are considered to belong to the same shot, and consequently a high similarity score (computed using the size of the maximum cardinality matching) should be encountered. Our search procedure uses dissimilarity measurements calculated between frames in a sliding window W. This sliding window is divided into two disjoint parts whose size is equal to r frames (see Fig. 1). More specifically, for the sliding window W of size $2r$, which is centered between frames f_k and f_{k+1}, we compute $2r$ lists of frame points $\mathrm{L}_{k-r+1}, \ldots, \mathrm{L}_k, \mathrm{L}_{k+1}, \ldots, \mathrm{L}_{k+r}$. Then, we generate point similarity graphs between lists of frame points which do not belong to the same part of the sliding window W (see Fig. 1). Finally, for a given sliding window W with radius r, the dissimilarity measure $\mathrm{DIS}^{\delta,\lambda}$ is calculated for each those graphs and used to compute the r-cumulative dissimilarity measure as follows.

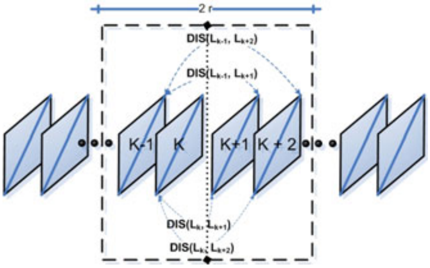


Fig. 1. Computation of cumulative dissimilarity for a sliding window W whose size is equal to 4 (i.e., with radius $r = 2$) and centered between frames f_k and f_{k+1}. The cumulative dissimilarity for frame f_k will be the summation of dissimilarity measures between frames from disjoint parts of the sliding window, i.e, $2\text{-}\mathrm{CDIS}_k^{\delta,\lambda} = \mathrm{DIS}^{\delta,\lambda}(\mathrm{L}_k, \mathrm{L}_{k+1}) + \mathrm{DIS}^{\delta,\lambda}(\mathrm{L}_k, \mathrm{L}_{k+2}) + \mathrm{DIS}^{\delta,\lambda}(\mathrm{L}_{k-1}, \mathrm{L}_{k+1}) + \mathrm{DIS}^{\delta,\lambda}(\mathrm{L}_{k-1}, \mathrm{L}_{k+2})$.

Definition 4 (r-Cumulative dissimilarity – $r\text{-}\mathrm{CDIS}_k^{\delta,\lambda}$**).** *Let* f_k *be the frame at location* k*,* $k \in [0, N-1]$ *and* L_k *be the list of frame points associated with that frame. So, for a* $2r$*-sized sliding window centered between frames* f_k *and* f_{k+1}*, the* r*-cumulative dissimilarity* $r\text{-}\mathrm{CDIS}_k^{\delta,\lambda}$ *can be calculated as*

$$r\text{-}\mathrm{CDIS}_k^{\delta,\lambda} = \sum_{i=k-r+1}^{k} \sum_{j=k+1}^{k+r} \mathrm{DIS}^{\delta,\lambda}(\mathrm{L}_i, \mathrm{L}_j). \tag{1}$$

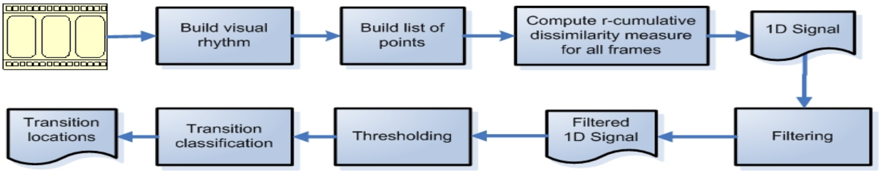


Fig. 2. Workflow for transition detection

Fig. 1 illustrates the computation of r-CDIS$_k^{\delta,\lambda}$. Finally, the transition detection problem can be stated as follows.

Definition 5 (Transition detection – TD). *The transition detection (TD) problem corresponds to the identification of all content changes on a video sequence. Thus, transition detection at any frame f_k can be defined as*

$$TD(\mathrm{V}_N, r, \lambda, \delta, \Delta) = \{k \in \mathbb{T} | f_k \in \mathrm{V}_N, r\text{-CDIS}_k^{\delta,\lambda} \geq \Delta\} \qquad (2)$$

where r-CDIS$_k^{\delta,\lambda}$ is the r-cumulative dissimilarity measure for a sliding windows W of radius r centered between frames f_k and f_{k+1}; and three specified thresholds λ, δ and Δ. λ corresponds to the maximum distance between two point locations, δ corresponds to the maximum dissimilarity allowed between two point values; and Δ corresponds to the minimum cumulative dissimilarity score needed to classify the location as transition.

One should notice that Δ may be either specified or an adaptive threshold can be used. To specify a single value for Δ that is best suitable for a given situation is not an easy task. Moreover, depending on parameter values, the transition detection approach stated by Equation 2 identifies all types of transitions.

3 Our Method for Transition Detection

As described before, the main goal of transition detection problem is to identify changes on a video sequence, such as cuts, dissolves, fades, and wipes, among others. In the proposed workflow, as described in Fig. 2, the first step of the process is the extraction of frame points from a visual rhythm [4] in order to construct lists of frame points for a specified window with $2r$ frames.

The main idea of our method is to compute the r-cumulative dissimilarity measure for a sliding window centered between two frames. It is important to remark that window size, i.e. $2r$ frames, where r is the radius parameter, is directly related to the gradual transition size that may be identified. For instance, let T be the length in frames of a transition. According to [8, 9, 11], for an ideal (gradual) transition (i.e. a linear dissolve between two almost still shots ends), the 1-dimensional signal obtained from sequence of dissimilarity values results in a *plateau* (an isosceles trapezoid) with width of $2r + T + 1$ (with a minor top base of $|2r - (T + 1)|$ frames and two slopes of $\min(2r, T + 1)$ frames) with maximum height for $2r = T + 1$ (when it degenerates into a triangle since the minor base of the trapezoid becomes a point). Therefore, if radius parameter is set to $(T + 1)/2$ a set of local maxima in the 1-dimensional dissimilarity signal will be associated with the location of transitions. Unfortunately, since the size of (gradual) transitions is not fixed our dissimilarity measure (calculated with a fixed radius) will produce a signal with high dynamics; and, consequently, a filtering step is needed to locate local maxima.

Once the r-cumulative dissimilarity measure $r\text{-CDIS}_k^{\delta,\lambda}$ is calculated for each frame at location k, the sequence of dissimilarity values is then stored as an 1-dimensional signal (see Fig. 3). Due to the hypothesis that the computation of the dissimilarity measure produces a local maximum with high dynamics into a 1-dimensional signal, morphological filters – (i) closing; and (ii) white top hat – can be applied in order to find or enhance these maxima (see Fig. 3).

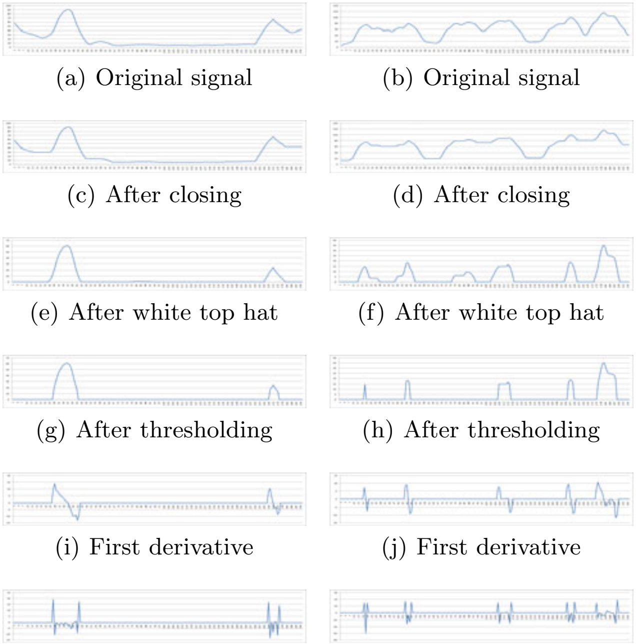

(a) Original signal
(b) Original signal
(c) After closing
(d) After closing
(e) After white top hat
(f) After white top hat
(g) After thresholding
(h) After thresholding
(i) First derivative
(j) First derivative
(k) Second derivative
(l) Second derivative

Fig. 3. Signal filtering, thresholding and classification of a transition: (left column) the first transition of this signal is gradual since $t_e - t_s \geq 3$; and (right column) the first transition of this signal is abrupt since $t_e - t_s \leq 2$

After the filtering step, a transition is associated with a local maximum that is larger than a specified threshold value (Δ), in order to prevent a great number of false positives related to effects, such as flashes and object/camera motions. Finally, the transition center is located where the first derivative of the 1-D signal changes sign (crosses zero) and the corresponding points have negative values of the second derivative. After identifying the maxima of the signal, we search

Table 1. Overall Values of F1 Score

Minimum cumulative dissimilarity ($\varDelta$)	F1 Score			
	Min	Max	Avg	Std Dev
05%	87.00%	91.45%	90.21%	0.94%
10%	88.09%	92.02%	90.25%	0.96%
15%	85.28%	90.94%	88.49%	1.13%

around the maximum for the start t_s and end t_e time instant of the transition (transition boundaries). Boundaries are detected as the points left/right of the maximum where the second derivative crosses zero in the so called inflection points. Since, a gradual transition has a certain duration, we consider that at least three frames should be involved to declare a gradual transition, i.e., $t_e - t_s \geq 3$ (see Fig. 3). If this condition does not hold, i.e., if $t_e - t_s \leq 2$ then an abrupt transition (cut) is declared (see Fig. 3).

4 Experiments

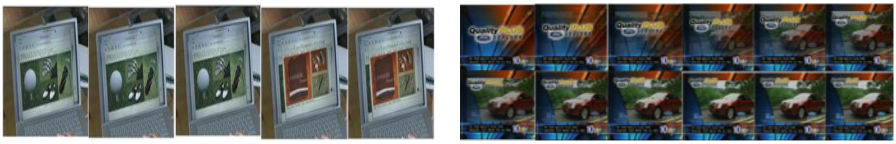
In our experiments, we have used two video datasets which presents a variety of different video genres. The first video dataset contains 20 videos – 1069 seconds (31796 frames) of MPEG-1 testing material with 570 transitions (47 cuts and 523 graduals). The second dataset contains 10 videos from TRECVID 2006 related to shot detection track, with 15160 seconds (467895 frames) of MPEG-1 testing material with 2509 transitions (1770 cuts and 759 graduals). In order to evaluate the results, we consider the precision, recall and F1 measure. According to [12], F1 is a combination of precision and recall and is maximized at the intersection of the two distributions. For that reason, F1 score is also called by [12] the best overall performance measure.

Several experiments, applied to the first dataset, have been conducted for 3 (three) different values of minimum dissimilarity score ($\varDelta$), i.e., for $\varDelta = 05\%, 10\%, 15\%$, for 3 (three) values of radius (r), i.e., $r = 09, 12, 15$, for 4 (four) values of maximum point dissimilarity (δ), i.e., $\delta = 05, 10, 15, 20$, and for 4 (four) values of maximum point distance allowed (λ), i.e., $\lambda = 05, 10, 15, 20$. Table 1 presents overall average values of F1 score obtained for those tests, together with minimum and maximum values and standard deviation for each value of minimum dissimilarity score ($\varDelta$). One can easily verify that best results found are associated with $\varDelta = 10\%$, while the set of tests with $\varDelta = 15\%$ produces the lowest overall average value of F1 score.

Table 2 presents a detailed view of these three quality measures (recall (R), precision (P), and F1 score) for the proposed method for several parameter settings (with $\varDelta = 10\%$). The proposed method achieves more than 92% recall with almost 92% precision (see Table 2 for window radius $r = 12$, $\delta = 10$ and $\lambda = 15$). Even the dataset is different for literature, our results are similar to (and even better than) the best results presented in [8] (their results are only

Table 2. Results for video dataset 1 and $\Delta = 10\%$

Window radius $r = 12$												
Point	Point dissimilarity (δ)											
distance	5			10			15			20		
(λ)	R	P	F1	R	P	F1	R	P	F1	R	P	F1
5	0.891	0.904	0.897	0.905	0.900	0.903	0.893	0.902	0.897	0.894	0.902	0.898
10	0.912	0.910	0.911	0.919	0.916	0.917	0.919	0.916	0.917	0.894	0.910	0.902
15	0.921	0.911	0.916	**0.924**	**0.916**	**0.920**	0.921	0.914	0.918	0.894	0.909	0.902
20	0.921	0.913	0.917	0.866	0.913	0.889	0.912	0.912	0.912	0.894	0.901	0.898



(a) False positive cut

(b) False positive gradual transition

Fig. 4. Examples of abrupt and gradual transitions that do not appear in groundtruth of TRECVID 2006

better for abrupt transitions), but our method uses a much simpler classification approach. Moreover, the number of gradual transitions in our first dataset is almost 92% (523 gradual transitons), while in [8] only 14% from the total number of transition are graduals (262 gradual transitions and 1571 cuts).

In our first dataset, since the transition average size is 23 frames, the best results (i.e., higher F1 scores) are associated with radius $r = 12$ $(= (23 + 1)/2)$ – as it should be expected. In our dataset, the shot average size is only 31 frames, so for larger values of r there is a great probability that two consecutive plateaus (i.e. transitions) merge into a single one. We consider the first dataset to tune the parameters. The proposed method, applied to the second dataset, achieves more than 75% recall with 68% precision and F1=71% for window radius $r = 12$, $\delta = 15$ and $\lambda = 15$). In order to understand the reasons behind this reduction in both recall and precision rates, we have to take a closer look into video content added to the dataset. The average size of gradual transitions is 15.4 frames, and consequently, the window radius must be $r = 8$ instead of $r = 12$ in order to decrease the dissimilarity measure and increase the precision rate. Also, as one reexamines TRECVID 2006 groundtruth, he might have some doubts about some false positives detected by our method. Fig. 4 presents a example of 2 video sequences that were detected by our method. The first one shows a “video-in-video” (in which there is a cut), which is a very hard problem to cope with during transition detection. The other is an example of effects that are very hard to classify (even for humans). One could claim that they are again examples of “video-in-video”, but they also could be taken as gradual transitions. Many teams of TRECVID 2006 have reported those same problems.

Also, the quality of video sequences is poor, that is very different of our first dataset in which there is no doubt about boundary classification. Unfortunately, the dissimilarity measure adopted does not allow our method to identify those effects.

5 Conclusion and Further Works

In this work, the size of the maximum cardinality matching calculated using a bipartite graph with respect to a sliding window is used as a dissimilarity measure in order to identify locations of abrupt and gradual transitions. The main contribution of our work is the application of a simple and efficient distance to solve a problem of video segmentation. According to experimental results, the performance of our method, when applied to the first dataset (more than 92% recall with almost 92% precision), is similar to (and even better than) the one proposed by [8] with lower computational cost since its classifications step is much simpler. In our experiments, we have used a “not so large” dataset – but it presents a huge number of gradual transitions (almost one transition for each two seconds), which makes the problem of abrupt and gradual transition detection even harder. However, transition detection results can be highly dependent on the testing material, which is usually scarce and not especially representative. So, as a future work, we plan to apply our approach to a large and representative video database in which the average shot size will be much greater than the size of the specified window. We also intend to investigate further strategies to cope with hard effects such as “video-in-video”.

Acknowledgments

The authors are grateful to PUC Minas (Pontificia Universidade Católica de Minas Gerais), MIC BH (Microsoft Innovation Center Belo Horizonte), CT-Info/MCT/CNPq (Project 551005/2007-6) and FAPEMIG (Project CEX PPM 126/08) for the financial support of this work.

References

1. Naphade, M.R., Mehrotra, R., Ferman, A.M., Warnick, J., Huang, T.S., Tekalp, A.M.: A high-performance shot boundary detection algorithm using multiple cues. In: Proc. Int. Conf. Image Processing - ICIP, pp. 884–887 (1998)
2. Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B.: A formal study of shot boundary detection. IEEE Transactions on Circuits and Systems for Video Technology 17(2), 168–186 (2007)
3. Wang, Y., Liu, Z., Huang, J.-C.: Multimedia content analysis. IEEE Signal Processing Magazine, 12–36 (2000)
4. Ngo, C.-W., Pong, T.-C., Chin, R.T.: Detection of gradual transitions through temporal slice analysis. In: CVPR, pp. 1036–1041. IEEE Computer Society, Los Alamitos (1999)

5. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying production effects. Multimedia Syst. 7(2), 119–128 (1999)
6. Fernando, W.A.C., Canagarajah, C.N., Bull, D.R.: Fade and dissolve detection in uncompressed and compressed video sequences. In: ICIP, vol. (3), pp. 299–303 (1999)
7. Zhang, H., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. Multimedia Syst. 1(1), 10–28 (1993)
8. Bescos, J., Cisneros, G., Martinez, J., Menendez, J., Cabrera, J.: A unified model for techniques on video-shot transition detection. IEEE Transactions on Multimedia 7(2), 293–307 (2005)
9. Grana, C., Cucchiara, R.: Linear transition detection as a unified shot detection approach. IEEE Transactions on Circuits and Systems for Video Technology 17(4), 483–489 (2007)
10. Guimarães, S.J.F., Patrocínio Jr., Z.K.G., de Paula, H.B.: A rotation and translation invariant algorithm for cut detection using bipartite graph matching. In: Proc. of the Tenth IEEE International Symposium on Multimedia (ISM 2008), pp. 104–110. IEEE Computer Society Press, Los Alamitos (2008)
11. Yeo, B.-L., Liu, B.: A unified approach to temporal segmentation of motion jpeg and mpeg compressed video. In: Proceedings of the International Conference on Multimedia Computing and Systems, pp. 81–88 (May 1995)
12. Whitehead, A., Bose, P., Laganiere, R.: Feature based cut detection with automatic threshold selection. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 410–418. Springer, Heidelberg (2004)