GRADUAL TRANSITION DETECTION USING COLOR COHERENCE AND OTHER CRITERIA IN A VIDEO SHOT META-SEGMENTATION FRAMEWORK

Efthymia Tsamoura, Vasileios Mezaris, Ioannis Kompatsiaris

Informatics and Telematics Institute / Centre for Research and Technology Hellas,
1st Km Thermi-Panorama Rd, Thessaloniki 57001, Greece
{tsamoura, bmezaris, ikom}@iti.gr

ABSTRACT

Shot segmentation provides the basis for almost all high-level video content analysis approaches, validating it as one of the major prerequisites for efficient video semantic analysis, indexing and retrieval. The successful detection of both gradual and abrupt transitions is necessary to this end. In this paper a new gradual transition detection algorithm is proposed, that is based on novel criteria such as color coherence change that exhibit less sensitivity to local or global motion than previously proposed ones. These criteria, each of which could serve as a standalone gradual transition detection approach, are then combined using a machine learning technique, to result in a meta-segmentation scheme. Besides significantly improved performance, advantage of the proposed scheme is that there is no need for threshold selection, as opposed to what would be the case if any of the proposed features were used by themselves and as is typically the case in the relevant literature. Performance evaluation and comparison with four other popular algorithms reveals the effectiveness of the proposed technique.

Index Terms— video shot segmentation, gradual transition, color coherence change, meta-segmentation

1. INTRODUCTION

Video manipulation and most notably retrieval has become in recent years an integral part of our everyday lives. This necessitates the development of techniques that cover the entire range of video analysis and understanding, to support efficient and effective indexing and retrieval of vast amount of video content. A basic analysis step contributing towards this goal is video shot segmentation.

A video shot is defined as a set of consecutive frames taken without interruption by a single camera. It it consequently a collection of frames with high degree of affinity that constitute a self-contained visual entity. In most cases shots contain fundamental information for events that occur in videos. Shot boundaries provide the basis for almost all high-level video content analysis approaches, validating it as one of the major prerequisites for successful indexing and retrieval in large video databases.

Transitions between video shots can be clustered into two main categories; abrupt transitions and gradual transitions. In abrupt transitions, the differences between two consecutive frames are significant. On the other hand, there are several kinds of gradual transitions including fades, dissolves and wipes, that are characterized by less evident differences between consecutive frames. The common characteristic between them is the existence of frames that clearly mark the transition period. Each frame in the transition period is typically a combination of the starting and ending frames.

In [1] transition detection is based on the analysis of intensity edges. An edge pixel that appears far from an existing edge pixel is defined as an entering pixel, while a previously existing edge pixel that disappears is defined as an exiting pixel. Transitions can be detected by counting the number of exiting and entering pixels and in particular the Edge Change Ratio (ECR). In [2] dissolve detection is implemented via analysis of a characteristic curve estimated from input video. Analytically, a dissolve modelling error is being compared with the actual dissolve modelling error estimated using input video. Let $\mathbf{I}_t$ and $\mathbf{I}_{t+\tau}$ be the starting and ending frames of a dissolve transition respectively. It is hypothesized that a dissolve is being constructed by a linear combination of the above frames. Under this assumption it was shown in [2] that the variance of pixel intensities exhibits parabolic shape during an actual dissolve. Candidate dissolve areas are identified using the first and second derivatives of the luminance variance curve and are verified using an adaptive threshold. A similar philosophy is exploited in [3]. Although the above modelling is reasonable, it is insufficient for detecting transition areas where pixel intensities change nonlinearly. C.W. Su et al. [4] proposed a model to overcome the above problem. A sliding window that covers a number of frames is used, in order to observe the change of pixel intensities. If the number of pixels that undergo monotonous intensity changes (either increasing or decreasing) within the window is above a threshold, they conclude that the frames included within the window belong to a dissolve area.

In this paper a novel approach for the detection of dissolve and fade transitions is presented. Two main directions are followed. As a first direction, novel efficient criteria such as color coherence change are proposed, exhibiting reduced sensitivity to motion activity. As a second direction, a method of combining them in a meta-segmentation scheme in order to achieve more accurate detection results, is proposed. The segmentation process under the proposed approach can be summarized as follows: selected features are initially computed for every video frame. Given a couple of consecutive frames, the distances between the above features are then computed forming $distance\ vectors$. These vectors are subsequently supplied to a trained binary classifier, whose output denotes the class membership of each of the examined frames, i.e. whether it is part of transition area or not.

The paper is organized as follows. In section 2, a set of new visual criteria showing less sensitivity to global or local motion effects are presented. In section 3, a machine learning approach based on a Support Vector Machine (SVM) classifier is employed for combining the aforementioned criteria for gradual transition detection. In section 4 experimental results and comparison with four popular gradual transition detection algorithms are reported and finally conclusions are drawn in section 5.

2. INDIVIDUAL CRITERIA FOR GRADUAL TRANSITION DETECTION

2.1. Color Coherence Change

Color descriptors have been widely employed for shot detection, with color histogram being employed most often. Although color histograms accurately describe color distributions, they provide no information on the spatial arrangement of colors in the image. To alleviate this drawback, the use of Color Coherence Vectors (CCV) has been proposed for applications that involve image retrieval [5]. Color coherence expresses the degree of color's accumulation in an image area. Coherent pixels belong to contiguous regions of size greater than ψ, in contrast to incoherent pixels. A color coherence vector represents this categorization of pixels in the image. In this work, the use of color coherence change as a criterion for gradual transition detection in video is proposed.

Before computing any coherence vectors, the color space needs to be quantized into N color classes. In this work, color quantization is performed using the Macbeth color pallet [6]. The Macbeth pallet consists of twenty four colors which were selected according to human color perception. Having 18 elementary and 6 gray level colors including white and black, it constitutes a good compromise between color diversity and color cluster compactness: colors can be described by 24 color clusters. The relatively limited number of clusters ensures robustness to slight color variations or noise effects. Pixel colors are mapped to one of the 24 colors of the Macbeth pallet, constructing a 24 bins histogram. Let $Z_i\ i = 1 \dots 24$ denote the Macbeth pallet color clusters. Then, pixel $\mathbf{p}(x, y) = [R_{xy}\ G_{xy}\ B_{xy}]$ is assigned to the color cluster Z_i for which the manhattan distance $d = manhattan([R_{xy}\ G_{xy}\ B_{xy}], Z_i)$ is minimized. We choose the $manhattan$ distance due to its better suitability for the RGB color space, as compared to other distance functions, e.g. the euclidian.

The next step is to classify the pixels of a given color class as either coherent or incoherent. As previously mentioned, a coherent pixel is part of a connected spatial region whose pixels belong to the same color class. A connected component C is a set of pixels such that for every couple of pixels $\mathbf{p}$ and $\mathbf{p}' \in C$, there is a path in C between them. For each Macbeth color cluster Z_i, $i = 1 \dots 24$, some of it's pixels will be coherent, while the others will be incoherent. Let c_i be the number of coherent pixels of Z_i and d_i the number of incoherent pixels. The total number of pixels belonging to Z_i is $c_i + d_i$, resulting in a Macbeth color histogram

$$\mathbf{M}_t = [c_1^t + d_1^t\ \ c_2^t + d_2^t \dots c_{24}^t + d_{24}^t] \tag{1}$$

The color coherence vector is then defined as:

$$\mathbf{G}_t = [(c_1^t, d_1^t)\ \ (c_2^t, d_2^t) \dots (c_{24}^t, d_{24}^t)] \tag{2}$$

In our experiments we set ψ, the size of the smallest coherent area to 1% of the number of pixels in each frame. The distance between frames $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ having $\mathbf{G}_t$ and $\mathbf{G}_{t-1}$ color coherence vectors respectively, is computed as:

$$D_t^G = \sum_{i=1}^{24} |(c_i^t - c_i^{t-1})| + |(d_i^t - d_i^{t-1})| \tag{3}$$

Computing distances between couples of consecutive frames based on color coherence vectors for an input video, a D_t^G curve, $t = 0, \dots, T$ is produced, where T corresponds to video duration. Values higher than a threshold indicate that the corresponding frames are identified as belonging to a transition area. A simple threshold estimation procedure, similarly to [4], can be followed to determine the required threshold value.

2.2. Macbeth Color Histogram Change

Color histograms have been previously employed for shot detection. In cases where the moving objects cover a small percentage of the image pixels, or the camera is not moving very fast, color distribution changes are relatively small since only a small portion of pixels are affected at each time. On the other hand, when a gradual transition occurs, values of color histogram bins may vary enough to detect transitions, since changes typically take place in a much bigger part of image. We have chosen to employ color histograms based on the Macbeth color pallet. The distance between frames $\mathbf{I}_t$ and

$\mathbf{I}_{t-1}$ using their Macbeth color histograms $\mathbf{M}_t$ and $\mathbf{M}_{t-1}$ estimated as previously described in section 2.1, can then be defined as:

$$D_t^M = \sum_{i=1}^{24} |(c_i^t + d_i^t) - (c_i^{t-1} + d_i^{t-1})| \quad (4)$$

Computing the distances between pairs of consecutive frames based on the Macbeth color histogram feature, a curve D_t^M, $t = 0, \ldots, T$ is produced. Higher D_t^M values indicate greater color differences between consecutive frames. A threshold can be set, similarly to the color coherence case for detecting the gradual transitions.

2.3. Luminance Center of Gravity Change

When a dissolve occurs, the spatial distribution of pixel intensities changes. A simple, yet neglected in the shot detection literature characteristic is the luminance center of gravity. It's definition is analogous to an object's center of mass: it is the point where luminance is concentrated on. Let $L_t(x, y)$ be the luminance image calculated for frame $\mathbf{I}_t$. Then the luminance center of gravity of the frame is computed as:

$$\mathbf{R} = [R_x \; R_y] \quad (5)$$

$$R_x = \frac{\sum_x x L_t(x, y)}{\sum_x L_t(x, y)} \quad (6)$$

$$R_y = \frac{\sum_y y L_t(x, y)}{\sum_y L_t(x, y)} \quad (7)$$

The distance of the luminance centers of gravity between frames $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$, having $\mathbf{R}_t$ and $\mathbf{R}_{t-1}$ respectively is

$$D_t^R = ||\mathbf{R}_t - \mathbf{R}_{t-1}||, \quad (8)$$

where $||.||$ denotes the euclidian distance. Computing distances between frames based on the luminance center of gravity feature for an input video, a D_t^R curve, $t = 0, \ldots, T$ is produced for which similarly to the previous criteria, a suitable threshold can be estimated.

A significant advantage of the above criterion is it's insensitivity to slight pixel luminance variations as well as to measurement noise effects. If a small percentage of pixels values changes due to local or global motion, then the location of the center of luminance is not likely to be significantly affected. On the other hand it is sensitive to dissolves, since in that case intensities would change for a significant amount of image pixels.

2.4. Monotonous Intensity Change

This is a criterion proposed in [4], where it is hypothesized that in a gradual transition, pixel intensities vary monotonously. The percentage of pixels with monotonously varying intensities is calculated and if it exceeds a certain threshold, a dissolve/fade transition is detected. Let $f(x, y, t) = L_{t+1}(x, y) - L_t(x, y)$. Then, the monotonous change of intensity is evaluated using the following equation:

$$g(x, y, t) = \begin{cases} 1, & f(x, y, t) f(x, y, t-1) \geq 0 \\ 0, & f(x, y, t) f(x, y, t-1) < 0 \end{cases} \quad (9)$$

Subsequently, the percentage of pixels with monotonously varying intensities at a given time t can be calculated as $D_t^I = \sum_{xy} g(x, y, t)$. In [4] a threshold is estimated for D_t^I, to indicate gradual transitions.

3. META SEGMENTATION APPROACH

Since a single criterion alone is difficult to accommodate for all possible effects that hinder gradual shot detection, a combination of multiple individual criteria can be employed to improve detection accuracy. To this end, a machine-learning classification approach is adopted in this work, based on SVM. Gradual transition detection, under the proposed SVM-based meta-segmentation technique, has two phases: training and evaluation. For training, the classifier must be supplied with a set of input vectors manually assigned to the appropriate class (transition or non-transition). Subsequently, at the evaluation phase new input vectors can be classified by the SVM classifier to one of the two learnt classes. In our case input vectors are made of the individual criteria defined in section 2, estimated from consecutive video frames. Jointly using all criteria presented in section 2 would result in a 4-dimensional distance vector $\mathbf{D}_t$ between frames $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$,

$$\mathbf{D}_t = [D_t^I \; D_t^M \; D_t^R \; D_t^G] \quad (10)$$

whereas using a subset of the criteria is also possible by defining a distance vector of lower dimensionality. For classification we used a C-SVM with a radial basis function kernel of 3^{rd} degree.

A significant advantage of the proposed scheme is that input vectors are automatically classified, without any threshold. The use of the SVM classifier bypasses the need for threshold selection by learning from the training set the optimal separating hyperplane. This is shown to boost the performance of gradual transition detection.

4. EXPERIMENTAL RESULTS

The performance of our algorithm was evaluated using two videos of the TRECVID-2007 video set. Both video sequences had a frame size of 352×288 pixels at 25 fps. Gradual transitions were evaluated against manually generated ground truth results. These videos exhibited a large amount of gradual transitions as well as intense motion events. It

must be emphasized that about 30% of the included gradual transitions were relatively easy to identify, since they did not involve significant motion or color changes. The remaining 70% involved global or local motion and/or luminance changes that made their automatic detection a challenging task. The use of such a challenging test set has made possible the comparative evaluation of the proposed approach under realistic conditions. A 5-minute segment of the above video set, including 9 gradual transitions, was used for training the proposed system, whereas the remaining 55-minute segments, which included 89 gradual transitions, were used in the evaluation experiments presented below. Precision and recall measures are employed for evaluating the performance of the proposed approaches. Table 1 shows the results of the proposed meta-segmentation approach as well as those using each individual criterion of the four proposed in section 2 alone. To evaluate the latter, the distance curves D_t^R, D_t^M, D_t^I and D_t^G were computed as described in section 2 and appropriate thresholds were estimated as in [4]. It should be noted that the results shown in Table 1 have been estimated considering gradual transitions only. These results are therefore not directly comparable with those of any evaluation activity considering both gradual and abrupt transitions at the same time, particularly since in the various test corpora gradual transitions typically account for only a small portion of the overall transitions.

Method	Recall	Precision
Standard Deviation of Pixel Intensities [2]	0.765	0.484
ECR [1]	0.654	0.125
EAG [3]	0.28	0.146
Monotonous Intensity Change (D_t^I) [4]	0.78	0.409
Luminance Center Of Gravity Change (D_t^R)	0.75	0.15
Macbeth Color Histogram Change (D_t^M)	0.78	0.45
Color Coherence Change (D_t^G)	**0.91**	0.595
Proposed meta-segmentation approach with $D_t = [D_t^I\ D_t^M\ D_t^R]$	0.78	0.67
Proposed meta-segmentation approach with $D_t = [D_t^I\ D_t^M\ D_t^R\ D_t^G]$	0.88	**0.73**

Table 1. Recall and precision evaluation of various gradual transition detection algorithms.

We compared our algorithm with four popular approaches for gradual transition detection: Edge Change Ratio (ECR) [1], Standard Deviation of Pixel Intensities [2], Effective Average Gradient (EAG) [3], and Monotonous Intensity Change [4]. Table 1 shows results of them for the employed test set. These results indicate that the individual gradual transition detection criteria proposed in this work are comparable or better than the ones previously proposed in [1]-[4]. The color coherence change criterion in particular is shown to significantly outperform the examined approaches of the literature. The combination of the various criteria in the proposed meta-segmentation approach is shown to further improve gradual transition detection accuracy. The contribution of the color coherence change criterion to the meta-segmentation approach is made evident by comparing the performance of the latter when using all four criteria ($D_t = [D_t^I\ D_t^M\ D_t^R\ D_t^G]$) to using the remaining three ones ($D_t = [D_t^I\ D_t^M\ D_t^R]$).

5. CONCLUSION

A new gradual transition detection scheme was proposed. New criteria were introduced showing less sensitivity to global or local motion effects. These criteria, each of which could serve as a standalone transition detection approach, were then combined using a machine learning technique, to result in a meta-segmentation scheme. We demonstrated that the latter, as well as some of the individual criteria proposed, outperform four popular approaches of the literature. Future work involves the integration in the meta-segmentation scheme of additional features that can contribute to further improvement in accurate gradual transition detection.

6. ACKNOWLEDGEMENT

The research leading to this paper was supported by the European Commission Framework Programme 6 under contracts FP6-045547 VIDI-Video, FP6-027685 MESH and FP6-027026 K-Space.

7. REFERENCES

[1] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying production effects," *Multimedia Systems*, vol. 7, no. 2, pp. 119–128, 1999.

[2] J.U. Won, Y.S. Chung, I.S. Kim, J.G. Choi, and K.H. Park, "Correlation based video-dissolve detection," *Proc. Int. Conf. on Information Technology Research and Education, (ITRE2003)*, pp. 104 – 107, August 2003.

[3] H.B. Lu, Y.G. Zhang, and Y.R. Yao, "Robust gradual scene change detection," *Proc. Int. Conf. on Image Processing, (ICIP 99)*, vol. 3, pp. 304 – 308, 1999.

[4] C.W. Su, H.Y.M. Liao, H.R. Tyan, K.C. Fan, and L.H. Chen, "A motion-tolerant dissolve detection algorithm," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1106–1112, 2005.

[5] Greg Pass, Ramin Zabih, and Justin Miller, "Comparing images using color coherence vectors," *Proc. ACM Int. Conf. on Multimedia, (MM 96)*, pp. 65–73, 1996.

[6] C. McCamy, H. Marcus, and J. Davidson, "A color rendition chart," *Journal of Applied Photographic Engineering*, vol. 2, no. 3, pp. 95–99, 1976.