Dissolve Detection in a video sequence based on Animate Vision*

M. Albanese
Università di Napoli
malbanes@unina.it

A. Chianese
Università di Napoli
angelo.chianese@unina.it

V. Moscato
Università di Napoli
vmoscato@unina.it

L. Sansone
Università di Napoli
sansone@unina.it

Abstract

The objective of video segmentation is to segment a video sequence into parts called shots corresponding to a continuous set of frames taken from one camera. Transitions between shots can be abrupt (*cuts*) or gradual. Abrupt transition can be easily detected while detection of gradual transition, such as dissolve, in still an unsolved problem. In this paper we present a novel approach for a reliable dissolve detection method based on *Animate Vision* theory.

1 Introduction

With the rapid progress in video technology, video has quickly become an essential component of todays multimedia applications, including VCR, video-on-demand, virtual walkthrough, etc.. Thank to the modern multimedia compression techniques, we have observed an exponential increase of digital videos.

The main feature of a video management system is the presence of an efficient indexing system to enable fast access to the stored data. This could be achieved by a set of semantic indices for meaningfully describing video scenes and a query capability for flexible specification and efficient retrieval.

Video segmentation is a fundamental process for automatic video analysis and detected shots can be used as a start point for the creation of a video summary and indexes for fast retrieval.

Shot changes can be divided in two categories: abrupt transitions and gradual transitions. Gradual transitions include camera movements: panning, tilting, zooming and video editing special effects: fade-in, fade-out, dissolve, wipe.

In particulary a dissolve occurs when one whole picture fades away while another whole picture appears. It provides a smooth restful transition, with its speed affecting the overall mood and flow of video sequences.

Dissolves are often used in dance and music pieces, in TV programmes and in some transitions in drama. It is also used in live sports to separate slow motion replays from the live action.

Abrupt transitions are very easy to detect as the two frames of consecutive shots are enough uncorrelated, moreover a considerable number of works has been reported. Gradual transitions are more difficult to detect as the difference between frames corresponding to two successive shots is reduced, so the segmentation process cannot be based on the assumption that similarity between consecutive frames is very high only if such frames belongs to the same shot. Due to large diffusion of such effects, the task of segmenting videos into shots can be very hard.

Let us consider a TV news network that needs to manage a large number of video data. Such videos are typically produced using a great number of transition to evolve from one news clip to another in a gradual manner. A new program may be produced using the several stored clips. In order to do that, a system for retrieving video segments starting from some information is required. Such a system supposes that video clips are indexed by associating

*This work has been carried out partially under the financial support of the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) in the framework of the FIRB Project "Middleware for advanced services over large-scale, wired-wireless distributed systems (WEB-MINDS)"

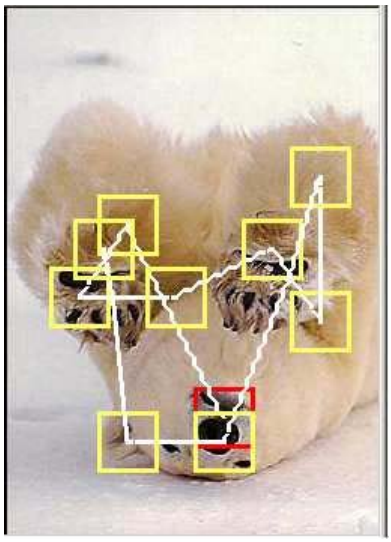

Figure 1: IP example

appropriate content information to each of the segments detected by a segmentation algorithm, which is able to recognize transition effects and to properly segment videos.

Among the numerous types of gradual transitions, dissolve is considered the most common one, but also the most difficult to detect. It is well known that an efficient dissolve detection algorithm which can be executed on a real video is still deficient and some authors consider dissolve detection as an unsolved problem.

To the best of our knowledge, a lot of algorithms have been proposed for fading (fade-in, fade-out) regions detection, while actually there are few works about the detection of other special effects as dissolve and wipe.

Lienhart [7] casts the problem of automatic dissolve detection as a pattern recognition and learning problem. Nam and Tewfik [9] use B-spline polynomial curve fitting technique to detect dissolve. In [6] Li and Wei proposed a dissolve detection method based on the analysis of Joint probability Images. In [12] Bul et al. proposed a dissolve detection algorithm based on statistical features of a image, while in [2] Liao el al. implemented a motion-tolerant algorithm for dissolve detection.

In this paper we present a simple and novel algorithm for dissolve detection in a video sequence based on Animate Vision.

The paper is organized as follows: in section 2 we recall the theoretical background on Animate Vision concepts and we describe as these concepts can be adapted to the problem of video segmentation; in section 3 we describe the dissolve detection algorithm; experiments and conclusions are presented and discussed in sections 4 and 5 respectively.

2 Animate Vision: a novel approach for video segmentation

In the most visual biological systems, only a small fraction of the information registered at any given time reaches levels of processing that directly influences behavior. The points inside of an image have not the same importance or salience, human eye attention is captured only from specified points, called *saliency point*.

A basic example is the generation of saccades, i.e. fast and repeated ocular movements that allow to acquire resolution image only at the most relevant part of the scene.

Animate Vision [1] is the visual biologic system capacity of quickly detecting interesting region of visual stimulus. The $Itti-Koch$ algorithm [4], based on $Koch-Ullman$ model [3, 5], allows the extraction of image saliency points.

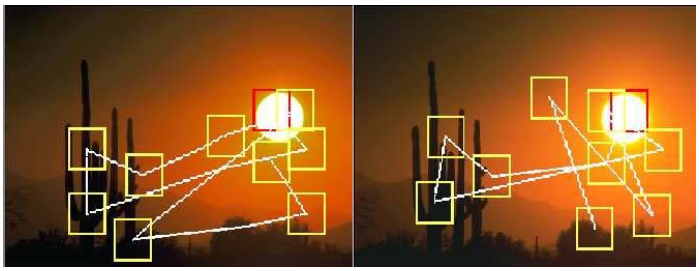

Figure 2: Similar Images

This process of attention selection is after followed by the definition of regions which surround saliency points; these regions are also called FOA ($Focus\ Of\ Attention$). So, with *Animate Vision* processing, an image can be represented from its *scanpath* (FOA sequence) [10]. Given two scanpaths extracted, respectively, from a target image and a test image, these can be compared in order to provide a similarity measure of the two images. To this end, each scanpath undergoes further processing/measurement, so that relevant information/features are derived: we denote the "flow" of such features, *Information-Path* (IP). In figure 1 an example of IP is shown.

In [8] we proposed a matching algorithm between two IPs, *Information-Path Matching algorithm*. It is based on the assumption, confirmed by experimental results, that similar images have similar IP as shown in figure 2.

Definition 2.1 (Image similarity) *Let I_1 and I_2 be two images, and $\mathcal{M}_{image}(I_1, I_2)$ the similarity measure between I_1 and I_2 returned by the IPM algorithm. I_1 and I_2 are similar iff*
$\mathcal{M}_{image}(I_1, I_2) > \xi$
being ξ a fixed threshold.

The threshold ξ is used as a lower bound to evaluate the similarity between two images.

Because a video is a frame sequence and a frame is an image, we can easily adapt IPM algorithm for abrupt transitions detection in a video sequence.

In a cut the last frame of previous shot is very dissimilar to the first frame of next shot, so they are not similar according definition2.1.

The real problem, in our approach, is the detection of gradual transitions as dissolve. During a dissolve, as the difference between consecutive frames is reduced, a frame is always similar to the next one, so the algorithm misses to detect the shot-change.

In section 3 we show how it is possible to detect this special kind of effect in a video, introducing a mathematical model of dissolve.

3 Dissolve Detection via Animate Vision

In this section we describe how our algorithm has been implemented using the *Animate Vision* concepts.

3.1 Mathematical Model for Dissolve

During a fade, a frame of a video sequence gradually appears or disappears. In particular fade-out occurs when the visual information gradually disappears, leaving a solid colour frame and fade-in occurs when the visual information gradually appears from a solid colour frame. Dissolve is a combination of fade-out and fade-in, superimposed on the same film strip.

Figure 3: Dissolve Effect

In other terms, the dissolve process is used to move on gradually from picture A to picture B. As the contribution of picture A changes from 100% to zero, the contribution of picture B changes from zero to 100%; if picture A is a solid colour, this process is a fade-in and, if picture B is solid colour, it is a fade-out.

Mathematically the dissolve effect can be modelled as in equation 1.

$$S_n(i,j) = \begin{cases} f_n(i,j) & 0 \leq n \leq L_1, \\ [1 - (\frac{n-L_1}{F})]f_n(i,j) + (\frac{n-L_1}{F})g_n(i,j) & L_1 \leq n \leq (L_1 + F), \\ g_n(i,j) & (L_1 + F) \leq n \leq L_2. \end{cases} \quad (1)$$

where $S_n(i,j)$ is the resultant video signal, $f_n(i,j)$ is the video signal corresponding to the first shot, $g_n(i,j)$ is the video signal corresponding to the second shot, L_1 is the length of $f_n(i,j)$, F is length of dissolve region, L_2 is length of the total sequence. All video signals are function of the coordinates (i,j) of a generic video frame point.

Given the above mathematical model, we expect that similarity function $\mathcal{M}_{image}$ presents a slow trend in such region. In particular we can observe that there exist a threshold δ such that:

$$|\mathcal{M}_{image}(S_n(i,j), S_{n+1}(i,j)) - \mathcal{M}_{image}(S_{n+1}(i,j), S_{n+2}(i,j))| < \delta \quad (2)$$

From the above considerations, we can base our approach on the study of the first derivative of $\mathcal{M}_{image}$ function.

3.2 Dissolve Detection Algorithm

Our dissolve detection algorithm is based on the computing of $\mathcal{M}_{image}$ function which uses the *Animate Vision* concepts, expressed in section 2, to return a similarity measure between two images.

In presence of dissolve effects (shown in figure 3), we observe that the $\mathcal{M}_{image}$ function has the characteristic behavior shown in figure 4.

The frame similarity measure decreases very slowly, like expected by the analysis of mathematical model, till a local minimum point (fade-out effect); then the function increases so slowly (fade-in effect).

It is easy to note, as shown in figure 5, that the first derivative function of $\mathcal{M}_{image}$ is constant and about zero in a frames region characterized by dissolving effects.

This feature is not enough to completely characterize a dissolve region: in fact we could have a similar trend of $\mathcal{M}_{image}$ function also in particular shots, e.g. a slow zoom. Another feature necessary to characterize a dissolve region is the "not similarity" between the edge frames. In other terms, the first and last frames of a dissolve region do not satisfy the property 2.1.

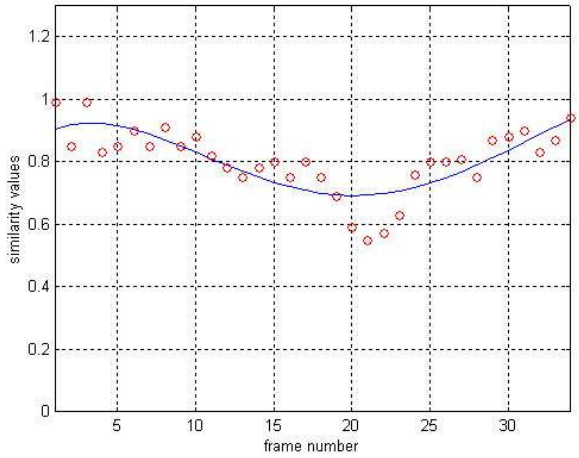

Figure 4: Similarity Function in a dissolve region

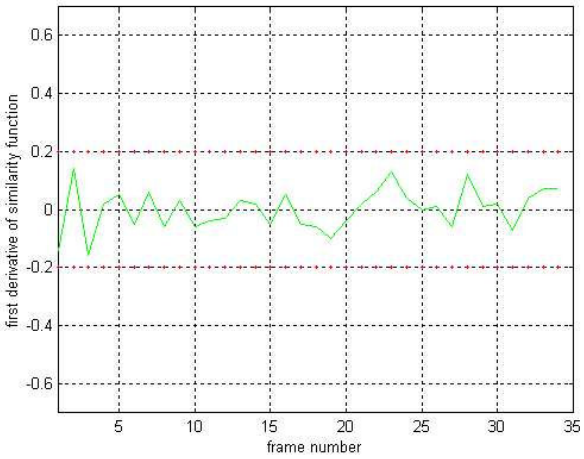

Figure 5: First derivative of frames similarity in presence of dissolving effects

Now we can introduce the following property:

Property 3.1 (Dissolve) *Given a sequence of frames $\mathcal{S}$ let f_{start} and f_{end} be respectively the first and last frame of such sequence. $\mathcal{S}$ is a dissolve region iff*

1. $-\delta \leq \mathcal{M}'_{image}() \leq \delta$ *in* $\mathcal{S}$
2. f_{start} *is not similar to* f_{end} *according definition* 2.1

This property is a direct consequence of the equation 1. In other terms , choosing a right video-sample, the similarity function, between consecutive frames in a dissolve region, has a well know and easy to detect trend.

Starting from the above considerations in the following we describe the *Animate Dissolve Detection Algorithm* ($ADDA$).

We denote with:

- I_{f_i} the Information Path extracted from the frame f_i
- vf_s and vf_e the vectors containing starting frames and ending frames of detected dissolve regions
- $\mathcal{M}_{image}(f_i, f_j)$ the similarity measure between frame f_i and f_j obtained by computation of IPM algorithm
- h the derivation step for image similarity function
- N the minimum number of $\mathcal{M}'_{image}$ consecutive zero values that characterizes a dissolve region

In a first stage, the first derivative of $\mathcal{M}_{image}$ is calculated. In a second stage frames belonging to a possible dissolve region are labelled on the base of the trend of $\mathcal{M}'_{image}$ function. In this the hypothetic dissolve regions are detected. Eventually the effective dissolve regions are built by property 3.1.

Algorithm 3.1 (Animate Dissolve Detection Algorithm ($ADDA$)) .

Compute the Information Path of every video frame I_{f_i}
Dissolve Condition verify and building of dissolve regions
$\mathcal{M}'_{image} = diff(\mathcal{M}_{image})/h$
$counter[1, ..., len(\mathcal{M}'_{image})] = 0$
$k = 0$
for $i = 1, ..., len(\mathcal{M}'_{image})$
  if $(\neg(0 - \delta \leq \mathcal{M}'_{image}(i) \leq 0 + \delta))$ **then**
    $k = i$
  else $counter[k+1] = counter[k+1] + 1$
  end if
end for
$f = 0$
for $j = 1, ..., len(\mathcal{M}'_{image})$
  if $((counter[j] > N)$ *and* $\neg(\mathcal{M}_{image}(f_j, f_{j+counter[j]-1}) > \xi))$
    $f = f + 1$
    $vf_s[f] = f_j$
    $vf_e[f] = f_{j+counter[j]-1}$
  end if
end for

The role of the constants δ and ξ is very important. The constant δ is used to verify if the first derivative function of $\mathcal{M}_{image}$ is constant and about zero in a given region, while threshold ξ is used to evaluate the similarity between the edge frames of an hypothetic dissolve region.

Video Sequence	Duration(frames)	# of dissolves
Movie 1 (The Patriot)	200000	47
Movie 2 (Dinosaurs)	92000	30
Documentary 1 (Desert Storm War Operation)	75000	26
Documentary 2 (Moscow)	82300	66
TV news (TG1)	114768	22
Total	564068	191

Table 1: Dataset

Video Sequence	Misssed Detection	False Allarms
Movie 1 (The life is Beautiful)	0%	7%
Movie 2 (Dinosaurs)	0%	3%
Documentary 1 (Desert Storm War Operation)	0%	6%
Documentary 2 (Moscow)	0%	9%
TV news (TG1)	0%	8%

Table 2: Experimental Results of dissolve detction

4 Experimental Results

To evaluate the performances of our dissolve detection algorithm, we have built a database of video sequences extracted from two documentaries, a TV news and two movies.

We captured video at 30 frames/s with 640*480 resolution. Our video database consisted of the sequences listed in table 1.

The video sequences are complex with extensive motion and a lot of graphical effects. The detection results are listed in table 2.

The algorithm's performances can be evaluated in terms of the number of not detected shot-changes MD(*Missed Detections*) and the number of detected wrong shot-changes FA (*False Allarms*), expressed as *recall* and *precision* [11].

$$recall = detects/(detects + MD) \tag{3}$$

$$precision = detects/(detects + FA) \tag{4}$$

We observed a 100% average recall rate and an average precision rate greater than 90%.

The main feature of our approach is the high recall. We have observed that each dissolve region is correctly detected on the base of the property 3.1, choosing right values of δ and N(in our experiments $\delta = 0.185$ and $N = 15$).

Recall and precision of our algorithm obviously depend also on the particulary threshold ξ chosen to verify similarity between the edge frames of a dissolve region. In our experiments, ξ is calculated as follow:

$$\xi = min_i(\mathcal{M}_{image}) \ \forall S_i \tag{5}$$

being S_i a possible dissolve region (only the first point of the property 3.1 is verified).

This choice of ξ allows to obtain a 100% recall at the expense of a variable precision but however high.

5 Discussion and final remarks

In this paper we have presented a novel approach for detection of cross dissolving in a video stream, based on Active Vision theory. Experimental results confirm the reliability of our approach and encourage to go on in this direction. Further research can be conduced in making the algorithm robust with respect to more type of transition. Furthermore research efforts can be addressed to extract motion information from video sequences in order to improve the performances of the segmentation algorithm.

References

[1] D. Ballard. *Animate Vision.* Artificial Intelligence n.48, 57-86, 1991.

[2] H. Y. Mark Liao L. H. Chen C. W. Su, H. R. Tyan. *A motion Tolerent Dissolve Detection Algorithm.* IEEE C2002, pp. 225-228, 2002.

[3] L. Itti and C. Koch. *Computational modelling of visual attention.* Natire Reviews-Neuroscience vol.2, 1-11, 2001.

[4] C. Koch and S. Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry.* Hum Neurobiol 4, 219-227, 1985.

[5] C. Koch L. Itti and E. Niebur. *A model of saliency based visual attention for rapid scene analysis.* IEEE Trans. on PAMI vol.20, 1254-1259, 1998.

[6] Z. N. Li and J.Wei. *Spatio-temporal Joint probability Images for Video Segmentation.* Proc. IEEE International Conference on Image Processing, pp.295-298, 2000.

[7] R. Lienhart. *Reliable Dissolve Detection.* Proc. SPIE:Storage and Retrieval for Media Databases, pp.219-230, 2001.

[8] V. Moscato A. Picariello M.Albanese, G. Boccignone. *Image Similarity based on Animate Vision: Information-Path Matching.* Multimedia Information System 8th Workshop, 66-75, 2002.

[9] J. Nam and A. H. Tewfik. *Dissolove Transition Detection Using B-Splines Interpolation.* IEEE International Conference on Multimaedia and Expo, 2000.

[10] D. Noton and L.Stark. *Scanpaths in the saccdice eye movements during pattern perception.* Visual Research (11), 929-942, 1990.

[11] R. Kasturi U. Gargi and S.H. Strayer. *Perfomance Characterization of Video-Shot-Change Dtections Methods.* IEEE Trans. on Circuits and systems for video technology vol.10, n.1, 1-13, 2000.

[12] D. R. Bull W. A. C. Fernando, C. N. Canagarajah. *Fade and Dissolve Detection Detection in Uncompressed and Compressed Video Sequences.* IEEE International Conference on Image Processing, pp.299-303, 1999.