Abstract

This paper deals with the construction of animation movie abstracts. The proposed video abstracts are computed using an adaptive selection of key frames obtained by color histogram analysis. The method exhibits three main parts. In the first step, a shot detection is performed. Then, within each shot, an adaptive number of key frames is selected. This selection is performed by analyzing cumulative distances between frame color histograms. In the third step, an inter-shot examination is performed. Using an iterative selection we provide a user-specified number of movie representative frames. The method has been tested on a database from the International Animated Film Festival that takes place yearly in Annecy.

1. Introduction

During the last years, increased processing power, bigger storage capacity and faster networks have induced a fast evolution in digital video. Consequently, this evolution has brought needs for new tools that aim at improving video archiving, cataloging, indexing and analyzing. In this context, video abstraction is a very important topic. It enables a quick browsing of a large video database and brings efficient content access and representation. In the literature, there are many works dealing with Video Abstraction [1]. Two basic forms of a video abstract can be distinguished:

- the moving-image abstract (or video skim) is the concatenation of a limited number of selected video segments,
- the still-image abstract (or static storyboard) is a set of key frames, constituting a salient collection of frames extracted from the video.

In this paper, we focus on still-image abstract. Many different techniques have been proposed for extracting key frames, from most simple ones consisting in using a uniform sampling of the video or selecting the central frame of each shot [2], to more sophisticated ones trying to catch the semantic structure of a video [3]. The main differences between these approaches come from the knowledge which is a priori required (number of key frames, shot boundaries, etc.) and from the frame features which are used (colors, regions, motion, interesting events).

Several mechanisms have been investigated to extract key frames. In [4], an unsupervised clustering scheme is proposed: the clustering space is a color space, and a key frame is the closest frame to the centroid of a big enough cluster. The computation of the dominant motion estimation can also be used [5]. In [6] authors used a multidimensional fuzzy histogram for each video frame based on a collection of appropriate features mixing the results provided by a color segmentation and a motion segmentation. In contrast to the above works which tend on maximizing the visual coverage of the key frame set, the methods proposed in [3] are based on the extraction of “interesting events”, key frames corresponding then to semantically important frames.

Whatever maybe the method, there is a fundamental question about the selection of a key frame: is it a frame which is very frequent in the sequence, this frequency being an argument in favour of its importance and representativeness? Or, on the contrary, is it a frame which is different from the other ones, being consequently a frame which probably catches the user attention? Generally, the methods mentioned above choose clearly one of these two strategies.
The method proposed here attempts to conciliate these two contrary objectives. It is a shot-based method. It means that a preliminary step is required to detect the shot boundaries. Then, once the shots are detected, the key frames are automatically selected in a new adaptive way. The frames features which are used are the color histograms. The key frame selection is performed in two levels: a first one dealing with each shot and a second one processing the key frames of the different shots.

Another specificity of the proposed method is its particular field of application which is the “International Animated Film Festival” [7] that takes place yearly (since 1960) in Annecy-France. Few work is dedicated to animation movies, most of the mentioned methods being applied to news, sport or documentary fields. Animation movies differ from other sequences in many respects: the events do not follow a natural way (objects or characters emerge and vanish without respecting any natural physical rules, the movements are not continuous), the characters can have any shape, a lot of visual effects are used and every animation movie has its own particular color distribution.

The rest of the paper is organized as follows: Section 2 gives the general description of the proposed approach. The intra-shot and inter-shot key frame selection are detailed in Section 3 and 4. Section 5 presents some experimental results. In conclusion, Section 6 contains final considerations.

2. General description

The proposed method is composed of three main steps and its general description is given in Fig. 1.

![Fig. 1: General description of the method](image)

The first step is the **shot detection**: the movie is divided into its fundamental video units called shots. A shot is a sequence of continuous frames and is obtained by detecting the video transitions (i.e., cuts, fades, dissolves). There is a large number of methods for video change detection in the literature [8]. In the following work, we use specially designed detection algorithms that we have developed to manage the difficulties created by the peculiarity of animation movies. The detection algorithms are presented in [9][10].

The second step is an **intra-shot key frame selection**: within each shot, an adaptive selection of frames is performed by analyzing color histogram distances between frames (Section 3). The obtained result is a first summary, composed of one or several relevant frames for each shot content.

The third and last step is an **inter-shot key frame selection** which is based on the analysis of a distance measure between the color histograms of the key frames detected in the previous step (see Section 4).

3. Intra-shot key frame selection

3.1. Preamble
As already discussed, the color information is a major feature in animation movies, thus the frames will be characterized by using color histograms.

3.1.1. Histogram computation
To get a fast processing, the histograms are computed in the following way:
- each frame is spatially sub-sampled by retaining 1 value for each 4x4 block,
- the colors are reduced by using a uniform quantification on the RGB components to get 125 colors.

Previous studies have evaluated the influence of these reductions. The proposed solution is a good compromise between a relevant characterization and a fast processing.

3.1.2. Frame distances
Key frame selection requires dissimilarity measures between frames. In the following, the distance between two frames $f_i, f_j$ is defined by the classical Manhattan distance, denoted $d_M(f_i, f_j)$, between their histograms. In the proposed approach, it is also necessary to evaluate the distance between a frame $f_i$ and a set of frames $\{f_j\}_{j=1}^{N}$. Inspired by the vector median filtering theory, we use the cumulative distance $d_i$, defined as:

$$d_i = \sum_{j=1}^{N} d_M(f_i, f_j)$$

$d_i$ is a measure of how $f_i$ is similar to the frames $\{f_j\}$.

3.2. One key frame selection
Within a shot, a key frame is supposed to be a good rep-
resenting of all the frames. According to this principle, the selection of only one key frame can be performed by selecting the frame which is the closest to all the other ones according to the distance $d_i$ proposed in equation 1. If $\{f_{i}\}_{i=1,N}$ denotes all the frames of the shot, the extracted key frame is $f_k$ where the argument $k$ is defined as:

$$k = \text{Argmin}_{i \in \{1,N\}} (d_i)$$

(2)

However, choosing a single key frame is not always satisfying, specifically when the shot comprises a certain variability brought by moving objects or camera motion. In this situation, a single key frame is not sufficient to represent the shot and a multiple key frame selection is necessary.

3.3. Multiple key frame selection

This multiple selection reveals two difficulties. The first one consists in detecting variability within a shot, and the second one is the selection of the key frames according to the different variability situations.

3.3.1. Detecting variability

Different solutions have been proposed in the literature like sufficient content change, maximum frame coverage or equal temporal variance. To keep on using color and to deal with the specificity of the animation movies, we propose a new strategy based on the analysis of the distance $d_i$ defined in equation 1.

For each shot, the histogram of the $d_i$ distances is computed. After observing and analyzing several examples of such histograms for a large dimension of animation movies, we have identified 4 main shapes given in Fig. 2:

- **shape 1**: all the $d_i$ cumulative distances are small. There is a low variability in the shot.
- **shape 2**: most of the $d_i$ cumulative distances are small, but there are a few frames which are very different from the others (large values of $d_i$ cumulative distances). It is a shot with many similar frames and a short period with variability (i.e. big moving objects).
- **shape 3**: the histogram is multimodal. There are different groups of similar frames. It corresponds to a shot composed of different static scenes connected by fast camera motions.
- **shape 4**: there is only one mode, but the $d_i$ cumulative distances have large values. Such a shot is composed of many different frames which may be caused by a continuous camera motion or by special visual effects.

These different shapes are classified by using a multimodal histogram analysis. Empirically chosen thresholds allow to distinguish the shape 1 and 4. In complement to this modal analysis, the following condition is used to identify shape 2:

$$\text{mean}(d_i) - \text{min}(d_i) < \frac{\text{max}(d_i) - \text{min}(d_i)}{C}$$

(3)

where $C$ is a constant value which has been fixed to 5 after different tests on a wide set of animation movies.

3.3.2. Key frame selection

According to each shape, a key frame selection is defined:
- **shape 1**: only one key frame is necessary in this situation. This key frame is selected according to eq.2.
- **shape 2**: two key frames are selected. The first one corresponds to the result of equation 2, and the second one is the frame corresponding to the maximum of $d_i$ distances.
- **shape 3**: a key frame is selected for each mode of the histogram. Within each mode, the key frame selection is obtained by using equation 2 applied on a limited set of frames of the considered mode.
- **shape 4**: this is the most difficult situation. The variability within the shot is important. The strategy used consists in selecting two key frames as for the shape 2: the frames corresponding to the result of equation 2 and to the maximum of $d_i$ distances.

The obtained abstract generally gives a large number of key frames, specifically for highly dynamic movies. It means that it is useful as a compact but complete representation for further analysis which is very interesting for movie analysis. For interactive tasks, a more compact abstract is necessary.

4. Inter-shot key frame selection

To get a more compact abstract, the proposed strategy consists in selecting a reduced set of very different
frames among the key frames given by the previous step.

Let us denote $S = \{f_k\}_{k=1}^K$, the set of key frames previously selected, and $K_0 (K_0 < K)$, the desired number of key frames ($K_0$ is chosen by the user in order to control the size of the abstract). The algorithm (Fig. 3) is an iterative one. At each iteration, we look for a new key frame far from the previously selected key frames.

**Algorithm:**

1. Select $f_n$ the first key frame so that $d_n = \min_{k} (d_k)$ // $k$ image index
2. $R \leftarrow R + \{f_n\}$ // $R$ = Abstract
3. Repeat
   - $S \leftarrow S - \{f_n\}$ // $S$ initial set
   - Select $f_{temp} \in S$ corresponding to the $\min_{f \in R} (d_{f}(f_n, f_{temp}))$
   - Select $f_n \in S$ corresponding to the $\max_{f \in R} (d_{f}(f_n, f_{temp}))$
   - $R \leftarrow R + \{f_n\}$
4. Until ($\text{Size}(R) = K_0$)

**Fig. 3 : Inter-shot key frame selection algorithm**

### 5. Experimental results

The proposed animation movie summarization method was tested on a set of 13 short animation movies from [7] with a total time of 74 min. The quality of the obtained abstracts has been evaluated by conducting a user study on a set of 30 persons. It provides a 90% satisfaction rate. Movie abstraction has somewhere a large part of subjectivity. So it is difficult to appreciate the effectiveness of the proposed method.

An example of several abstracts obtained for some shots of the movies: “Ferrailles”, “Le Roman de Mon Ame” and “The Buddy System” is illustrated Fig. 4.

**Fig. 4 : Shot abstraction: multiple key frame selection**

This abstract provides us with a complete content representation of the movie and could be used for further analysis such as movie color statistics [10].

The inter-shot key frame extraction gives a compact representation of the movie well suited for movie database browsing applications. An example of a compact abstract for the movie “Le Roman de Mon Ame” is presented in Fig. 5.

**Fig. 5 : 10 frame abstract (“Le Roman de Mon Ame”)**

Acknowledgments: authors would like to thank Citia and Folimage company which provides the database.

### 6. Conclusion

In this paper we have proposed a new method to define two still image movie abstracts. The first abstract is obtained in an adaptive way by analyzing the distances between the color histogram of the different frames and by identifying specific shot schemes. The second image abstract is a compact representation which gathers both the most frequent and the most different frames in the sequence. Several tests have confirmed the global quality of these abstracts.

### References