Key frame selection to represent a video

Frédéric Dufaux
Compaq Computer Corp., Cambridge Research Lab.,
dufaux@crl.dec.com

Abstract

This paper describes a technique to automatically extract a single key frame from a video sequence. The technique is designed for a system to search video on the World Wide Web. For each video returned by a query, a thumbnail image that illustrates its content is displayed to summarize the results.

The proposed technique is composed of three steps: shot boundaries detection, shot selection, and key frame extraction within the selected shot. The shot and key frame are selected based on measures of motion and spatial activity and the likeliness to include people. The latter is determined by skin-color detection and face detection.

Simulation results on a large set of video from the Internet, including movie trailers, sports, news, and animation, show the efficiency of the method. Furthermore, this is achieved at a very low complexity cost.

1. Introduction

This paper introduces a technique to automatically extract from a video sequence a single key frame representative of its content. The technique is most useful in a system to search video on the World Wide Web [1] such as the Altavista Video Search [2]. In this context, the result of a query is summarized by displaying a thumbnail image for each returned video. The technique can straightforwardly be extended to extract n key frames. In this case, a succession of n thumbnail images, for instance combined in an animated GIF, can be displayed for each returned video.

Because of its intended use, the technique should satisfy a number of constraints. First, it should be generic and make no assumptions on the video. Second, it should be robust and efficient to handle the low quality (highly compressed and low resolution) video that can be find on the Internet. Finally, it should be fast to economically process a very large number of video sequences.

A number of techniques have been published in the literature to extract key frames from a video sequence. However, these techniques address the problem of extracting all the key frames of a video with the goal of producing a storyboard representation for video browsing. Basically, these techniques first divide the sequence into shorter temporal units named shots. A shot is defined as a sequence of frames captured by a single continuous operation of the camera. Existing techniques for shot detection mainly relies on measures of frame-to-frame change [3]. Once shot detection has been completed, key frames are extracted for each shot. For instance, the first frame of each shot can be selected as key frame. If significant changes (e.g. color or motion) occur within a shot, more than one key frame can be selected for this shot. In [4], a clustering technique is used for this purpose.

In contrast, our problem consists in extracting a single key frame representative of a video. For this purpose, we have developed a technique composed of three steps. First, the video is segmented into shots, then a shot is selected, and finally a key frame is extracted from the selected shot. To select a key frame representative of a video is a highly subjective task. In our approach, we combine several measures to produce an interesting and informative key frame which tends to give a good idea of what a video is about. More specifically, the process to select the shot and key frame relies on measures of motion activity, spatial activity, skin-color pixels [5] and face detection [6]. By taking into account skin-color and faces, we increase the likeliness of the selected key frame to include people and portraits, such as close-up of movie actors, and therefore to produce interesting key frames.

Combining several measures and making no assumptions on the video content results in a robust and generic procedure as shown in simulation results on a large set of video from the World Wide Web. Furthermore, as the process relies on measures easily computed from the video, the computational complexity is kept low.

2. Proposed technique

The proposed technique, which produces a single key frame, is composed of three steps. First, the video is divided into shots, then the best shot is determined, and finally a representative frame is extracted from the selected shot. Clearly, the technique can be extended to extract n key frames by selecting n shots and subsequently a key frame from each selected shot.

In order to speed up the process, spatial and/or temporal subsampling can be applied on the incoming

0-7803-6297-7/00/$10.00 © 2000 IEEE

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on January 13,2025 at 08:37:07 UTC from IEEE Xplore. Restrictions apply.

video prior to extracting the key frame. Furthermore, on long video sequences, the process can be restricted to the beginning of the video.

2.1. Shot detection

Shot boundaries are identified by combining the detection of discontinuities in motion activity and changes in pixel value histogram distribution.

Motion activity is simply measured by pixel-wise frame difference

$$SAD(k) = \sum_{i,j} \left| I(i,j,k) - I(i,j,k-1) \right| ,$$

where $I(i,j,k)$ denotes the image luminance at pixel location (i,j) and frame k. Note that the *SAD* computation could use motion compensation. However, this results in very little gains at a significantly higher computational cost.

Histogram distribution dissimilarity is measured by the Kolmogorov-Smirnov statistical test [7] applied on the histograms of two consecutive frames

$$P_{KS}(k) = Q_{KS}(\sqrt{N/2} * D(k)) \text{ with}$$

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{(j-1)} e^{-2j^2\lambda^2} \text{ and}$$

$$D(k) = \max_x (CD(x,k) - CD(x,k-1)),$$

where $CD(x,k)$ is the cumulative distribution of the luminance histogram for frame k and x denotes the grayscale values (for a 256-level grayscale luminance $x \in [0,255]$). $P_{KS}(k)$, which satisfies $P_{KS}(k) \in [0,1]$, is the probability of the hypothesis that the distributions of luminance histograms of frames k and k-1 are the same.

2.2. Shot selection

The best shot is selected based on high motion and spatial activity and the likeliness to include people. Furthermore, long shots are given a preference, as they often show the intent of the content producer to emphasize a portion of the video.

Motion activity is again measured by frame difference $SAD(k)$. Spatial activity is computed by the entropy of pixel values distribution

$$H(k) = -\sum_x p(x,k) \log_2 (p(x,k)),$$

where $p(x,k)$ is the probability of the grayscale value x in the luminance histogram of frame k. For a 256-level grayscale luminance, $H(k) \in [0,8]$.

The presence of humans is determined by skin-color detection and face detection. The skin-color detector [5] is a learning-based system trained on large amounts of labeled data sampled from the World Wide Web. It returns for each frame the percentage of pixels classified as skin, denoted $S(k)$, $S(k) \in [0,1]$.

The face detection is performed using the neural network technique described in [6]. The face detector returns the number of faces in each frame, denoted $F(k)$, $F(k)$=0,1,2.... The technique in [6] was mainly developed for and applied to photographs. As frames from Internet video have typically a much lower quality (low resolution, high compression) than still images, applying the face detector to video tends to significantly decrease its performance.

Consequently, the face detector is prone to false-negatives and false-positives. False-negatives are mainly due to rotated, small, or occluded faces. Fortunately this is not too detrimental to the key frame extraction performance. However, the issue of false-positives cannot be so easily dismissed. Indeed, the performance of the key frame selection directly depends on a low number of false-positives. In order to alleviate this problem, a tracking system is introduced. Namely only faces tracked across several consecutive frames are retained. This significantly decreases the number of false-positives.

The neural network face detector is applied on a block of 20 by 20 pixels [6]. In order to detect faces at various scales and positions, a low-pass pyramid is build from the input image and the various scales are scanned with a 20x20 pixels window as described in [6]. Pyramid scaling parameter along with scales and positions to apply the face detector can be automatically determined based on the spatial resolution of the video sequence and computational complexity constraints.

Denoting k_i and k_j the first and last frame of a shot respectively, we define

$$MEDSAD = \text{MED}(SAD(k_i),...,SAD(k_j)),$$

$$MEDH = \text{MED}(H(k_i),...,H(k_j)),$$

$$MEDS = \text{MED}(S(k_i),...,S(k_j)),$$

$$SUMF = \sum_{k=k_i,...,k_j} F(k).$$

where MED() denotes the median operator. For each shot, a score combining the different measures is computed as

$$\text{Score(shot)} = w_{SAD} \frac{MEDSAD}{\sigma_{SAD}} + w_H \frac{MEDH}{\sigma_H} + w_S \frac{MEDS}{\sigma_S} + w_F \frac{SUMF}{\sigma_F} + w_T \frac{T}{\sigma_T}$$

where T is the length of the shot expressed in seconds, σ_{SAD}, σ_H, σ_S, σ_F and σ_T are the standard deviations of *MEDSAD*, *MEDH*, *MEDS*, *SUMF* and T respectively computed on a large data set, and w_{SAD}, w_H, w_S, w_F and w_T

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on January 13,2025 at 08:37:07 UTC from IEEE Xplore. Restrictions apply.

are weighting factors which can be adjusted heuristically. Finally, the shot with the highest score is selected

2.3. Key frame selection

To complete the process, the key frame is chosen from the selected shot based on a low motion activity (to avoid blurring or excessive coding artifacts), a high spatial activity, and the likeliness to include people. For each frame within the selected shot, a score combining the different measures is computed as

$$\text{Score(frame)} = w_H \frac{H(k)}{\sigma_H} + w_S \frac{S(k)}{\sigma_S} + w_F \frac{F(k)}{\sigma_F} - w_{SAD} \frac{SAD(k)}{\sigma_{SAD}}$$

and the frame with the highest score is selected.

3. Simulation Results

Experimental tests have been carried out on a set of 893 video sequences sampled from the World Wide Web, including movie trailers, sports, news, and animation.

A subjective evaluation has been performed to evaluate the quality of the selected key frames. To compare our results to a reference, we used a very simple algorithm which consists in selecting the middle frame of a video sequence. Note that choosing the middle frame is significantly better than choosing the first frame. Results were rated using three levels of quality:

- poor corresponding to semantically meaningless frame, for instance completely black frame,
- fair corresponding to semantically meaningful frame of low visual quality because of blur, blending,...,
- good corresponding to semantically meaningful frame of high visual quality.

Results of this subjective evaluation are reported in Table 1.

	middle frame	proposed technique
poor	7	0
fair	38	27
good	848	866

Table 1: subjective evaluation of key frames obtained by the middle frame of a sequence and by our proposed technique.

The key frames resulting from our technique were rated as good in 97% of cases, while the remaining 3% were considered as fair, therefore outperforming the middle frame algorithm. Despite its simplicity, the middle frame technique is surprisingly efficient. However, it results in poor key frames in about 1% of cases.

Figure 1 shows the motion activity $SAD(k)$ and histogram dissimilarity $1/P_{KS}(k)$ for a movie trailer. Peaks indicate shot boundaries. By combining both measures, high shot detection accuracy is achieved.

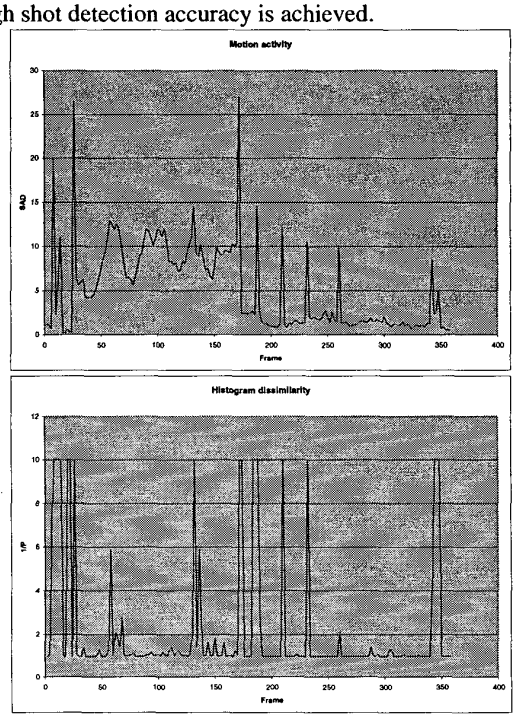

Figure 1: Motion activity ($SAD(k)$) and histogram dissimilarity ($1/P_{KS}(k)$).

Figure 2 shows a sample frame, the pixels classified as skin-color, and the detected face.

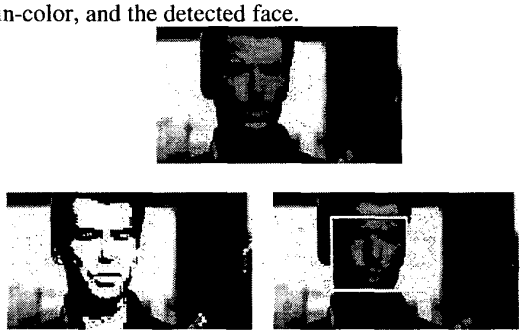

Figure 2: Example of a frame, skin-color pixels (in white), and detected face (in white square).

On a set of 30 movie trailers, the face detection results on average in a false-positives rate of approximately 30%. Despite this relatively low performance, the face detection is still very useful in the key frame selection process. To illustrate this point, Figure 3 shows four examples of key frame extracted from movie trailers with and without face

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on January 13,2025 at 08:37:07 UTC from IEEE Xplore. Restrictions apply.

detection. Clearly, the face detection increases the occurrence of portraits and hence results in more meaningful key frames.

Figure 3: Examples of key frame (left without face detection, right with face detection) for the trailers 'The Fifth Element', 'Les Misérables', 'Saving Private Ryan', 'Tomorrow Never Dies'.

In terms of complexity, the algorithm runs in real-time on a 500 MHz Pentium II. Note however that the face detection results in a significant complexity increased. Indeed, the system runs 10 times faster than real-time when not using the face detection.

4. Conclusions

This paper describes a technique to extract a single key frame from a video sequence. By combining measures of motion activity, spatial activity, skin-color pixels and face detection, the proposed technique results in an efficient and robust procedure. Its efficiency has been demonstrated in simulation results on a large set of video from the World Wide Web. Finally, the computational complexity is kept low as these measures are easily computed from the video sequence.

5. References

[1] M.J. Swain, "Searching for Multimedia on the World Wide Web", TR 99-1, CRL, Compaq Computer Corp., March 1999 (http://www.crl.research.digital.com/publications).

[2] Altavista Video Search (http://www.altavista.com).

[3] B.L. Yeo, B. Liu, "Rapid Scene Analysis on Compressed Video", IEEE Trans. on CSVT, **5** (6): 533-544, 1995.

[4] Y. Zhuang, Y. Rui, T.S. Huang, S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering", Proc. of Int. Conf. on Image Proc., Chicago, Oct. 1998.

[5] M.J. Jones and J.M. Rehg, "Statistical Color Models with Applications to Skin Detection", TR 98-11, CRL, Compaq Computer Corp., Dec. 1998 (http://www.crl.research.digital.com/publications).

[6] H.A. Rowley, S. Baluja, T. Kanade, "Neural Network-Based Face Detection", IEEE Trans. on PAMI, **20** (1):23-38, 1998.

[7] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, "Numerical Recipes in C, The Art of Scientific Computing", Cambridge University Press, 1988.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on January 13,2025 at 08:37:07 UTC from IEEE Xplore. Restrictions apply.