Key-frame extraction based on motion acceleration

Yanzhuo Ma, Yilin Chang, and Hui Yuan
Xidian University, National Key Laboratory of ISN, P.O. Box 108, Xi'an 710071, China
E-mail: yzma@mail.xidian.edu.cn

Abstract. Building on the argument that the change of motion states attracts more attention than the motion itself, this letter develops a novel method for key-frame extraction based on motion acceleration vectors. Different from the traditional methods using maximal or minimal motion energy, the proposed method uses the change of motion states, in magnitude and phase, of the main moving objects as the metric for key-frame extraction. Experimental results show that although similar objective performance is achieved by using the proposed method to that achieved with a widely used method based on motion energy, the key frames extracted by the proposed method are more consistent with human perception.
[DOI: 10.1117/1.2977795]

Subject terms: key-frame extraction; motion acceleration; main moving objects.

Paper 080336LR received Apr. 30, 2008; revised manuscript received Jul. 20, 2008; accepted for publication Jul. 30, 2008; published online Sep. 16, 2008.

1 Introduction

Among existing methods for key-frame extraction, those based on motion energy (ME) have proved to be both effective and computationally efficient.[1–4] This kind of methods is based on the simple idea that the more motion in the scene, the more interest of people should be attracted. Accordingly, local maximal or minimal ME, related to the motion magnitude, is usually employed as the metric for key-frame extraction. However, the extracted key-frames using an ME-based method are not representative in that many motions exist in most frames of video sequences. Naturally, objects keep normal motion states of rest or of uniform motion in a straight line unless compelled by external forces to change that state. The change of motion states is unpredictable and therefore more interesting to people than the motion itself. Therefore, the frames with an object changing its motion states, such as start, stop, acceleration, deceleration, or direction change, will provide more information and attract more attention than the frames containing uniform motion scenes. Thus, extracting key frames based on the changes of motion states, which can be uniformly represented as acceleration of the moving objects, is more consistent with human perception.

Motivated by this consideration, we define the frames with the most significant acceleration (MSA) of the main moving object as key frames, and accordingly propose a novel key-frame extraction method. The key frames obtained by the proposed method can reflect changes of motion states of the main moving objects such as moving in, moving out, and starting to change the motion direction or amplitude.

2 Proposed Method

In this letter, key frames are defined as the frames with the MSA. Generally, acceleration $\mathbf{a}(t)$ is defined as

$$\mathbf{a}(t)=\frac{\mathrm{d}\mathbf{v}(t)}{\mathrm{d}t}=\frac{\mathrm{d}\mathbf{v}_x(t)}{\mathrm{d}t}+\frac{\mathrm{d}\mathbf{v}_y(t)}{\mathrm{d}t}=\mathbf{a}_x(t)+\mathbf{a}_y(t), \tag{1}$$

where $(\mathbf{v}_x(t), \mathbf{v}_y(t))$ are the horizontal and vertical components of the velocity $\mathbf{v}(t)$, respectively. Let $\mathbf{v}(t-1)$ and $\mathbf{v}(t)$ denote the motion vectors (MVs) of a moving object at times $t-1$ and t, respectively. Then the acceleration vector $\mathbf{a}(t)$ of the object can be expressed as

$$\begin{aligned}\mathbf{a}(t)&=\mathbf{v}(t)-\mathbf{v}(t-1)\\&=[\mathbf{v}_x(t)-\mathbf{v}_x(t-1)]+[\mathbf{v}_y(t)-\mathbf{v}_y(t-1)]=\mathbf{a}_x(t)+\mathbf{a}_y(t).\end{aligned} \tag{2}$$

In each frame of video sequences, it is often found that many blocks have the same MV. If the greatest number of blocks corresponding to a nonzero MV (denoted as $N_{\max}$) is greater than a typical threshold T_N (usually expressed as a percentage, 2% as an example, of the number of the blocks in a frame), then these blocks are considered as belonging to the main moving object, and their MV is defined as the feature MV (FMV) of the frame. Otherwise, if $N_{\max}$ is less than T_N, the FMV will be set to zero. The framework of the proposed method is shown in Fig. 1. The acceleration vector of the main moving object, denoted as $\mathbf{a}_m(t)$, can therefore be computed as

$$\begin{aligned}\mathbf{a}_m(t)&=[\mathbf{mv}_x(t)-\mathbf{mv}_x(t-1)]+[\mathbf{mv}_y(t)-\mathbf{mv}_y(t-1)]\\&=\mathbf{a}_{mx}(t)+\mathbf{a}_{my}(t),\end{aligned} \tag{3}$$

where $(\mathbf{mv}_x, \mathbf{mv}_y)$ are the components of the FMV. The vector $\mathbf{a}_{\mathrm{m}}(t)$ in Eq. (3) can also be presented as

$$\mathbf{a}_m(t)=|\mathbf{a}_m(t)|\cdot\exp[-j\theta(t)], \tag{4}$$

where $|\mathbf{a}_m(t)|$ and $\theta_m(t)$ represent the magnitude and angle of the acceleration vector $\mathbf{a}_m(t)$, respectively.

The frames with typically high values of $|\mathbf{a}_m(t)|$ should be considered as candidates of key frames. However, the changes of motion direction of an object should also be considered, due to their importance to the human perception. Therefore, $|\mathbf{a}_m(t)|$ is weighted by a factor $w(t)$ to construct $a_w(t)$ as

$$a_w(t)=w(t)|\mathbf{a}_m(t)|. \tag{5}$$

For simplicity, $w(t)$ is defined as

0091-3286/2008/$25.00 © 2008 SPIE

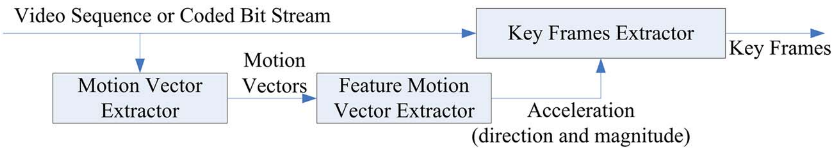

Video Sequence or Coded Bit Stream
Key Frames Extractor
Key Frames
Motion Vectors
Motion Vector Extractor
Feature Motion Vector Extractor
Acceleration
(direction and magnitude)

Fig. 1 Key-frame extraction based on acceleration.

$$w(t) = \begin{cases} 4 & \text{if the FMV direction reverses either} \\ & \text{horizontally or vertically at time } t, \\ 2 & \text{else if changing from still to moving} \\ & \text{or moving to still,} \\ 1 & \text{else (no direction change).} \end{cases}$$

Now the frames with typically high values of $a_w(t)$ are considered as candidates.

A motion change may last for several frames. To determine which frame is the most important one, a new function $a_{c,m}(t)$, which is a convolution of $a_w(t)$ with a time window, is introduced as

$$a_{c,m}(t) = \frac{1}{W} \sum_{\Delta t=-(W-1)/2}^{(W-1)/2} a_w(t+\Delta t), \tag{6}$$

where W is the width of the time window. We choose an odd value for W due to the symmetry of the time window with respect to its center t. Then the frames corresponding to the peaks (i.e., local maxima) of $a_{c,m}(t)$ will be extracted as key frames. It should be noted that a proper window size W can be selected according to the time resolution of video sequences. For a sequence with a high frame rate and/or slow motion in content, a large W should be selected.

3 Experimental Results

Four test sequences, namely "Erik," "Football," "Claire," and "Foreman," were employed in the experiments. For performance evaluation, the proposed key-frame extraction method is compared with the ME-based method presented in Ref. [2] by extracting the same number of key frames. Both objective performance as measured by shot reconstruction degree (SRD) and subjective performance are compared in this letter.

The SRD is the average peak SNR of the interpolated frames, based on inertia with respect to their original frames in the sequence.[2] Figure 2 shows the curves of the mean SRDs of the four sequences in quarter common intermediate format (QCIF) by the two compared methods with different extracted-key-frame ratios from 2% to 12%.

It is observed from Fig. 2 that the two methods achieve similar SRD performance. For the cases where the percentage of key frames is below 6%, the proposed method outperforms the ME-based method[2] by about 0.1 to 0.3 dB. For larger percentage of key frames, the ME-based method achieves slightly better performance, by up to 0.2 dB.

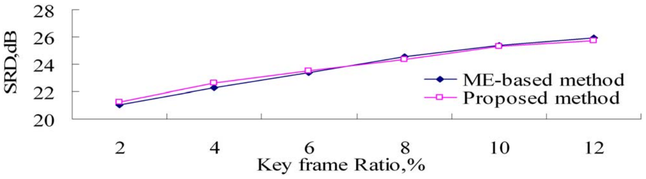


Fig. 2 SRD of proposed method and ME method.

Although the overall difference in SRD performance between the two methods is very small, the subjective performance of the proposed method is better. Specifically, subjective results show that the frames in which changes of object (the head in "Claire") movement or scene changes (in "Football" and "Foreman") happened can be well extracted from all the three sequences by the proposed method, but not always by the ME-based method. Due to limitations of space, only the key frames extracted from "Erik" in QCIF and "Foreman" in CIF are shown in Figs. 3 and 4, respectively.

A detailed analysis of the results of the "Erik" sequence, which is representative of typical alternations of uniform motion and motion changes, is given as follows as an example. Figure 3(a) shows the track of the nose position changing horizontally with time. The whole track can be approximately recovered from several inflection points, viz., at frames 1, 14, 27, 39, and 50. Therefore, these frames are the key frames used as the benchmark frames to evaluate the performance of the key-frame extraction methods. The same number of key frames extracted by the ME-

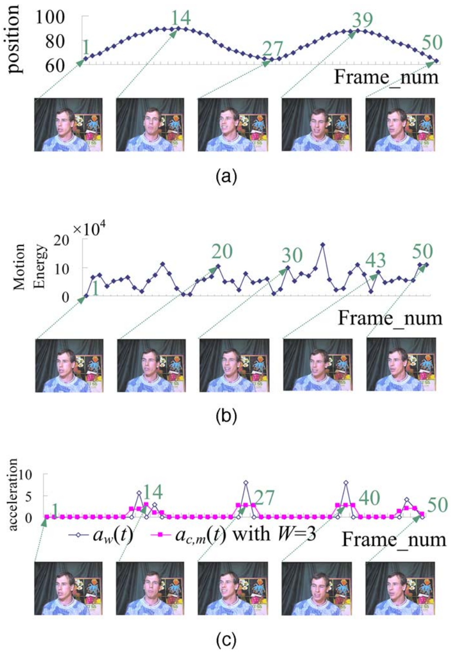
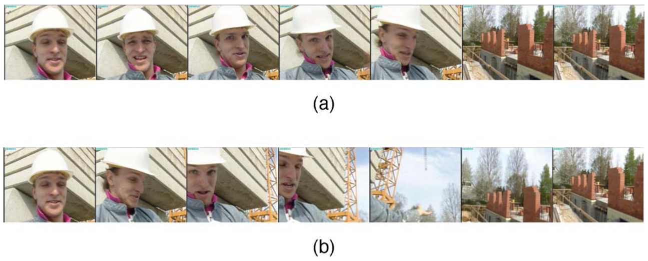


Fig. 3 Extracted key frames ("Erik," QCIF, key-frame ratio 8%): (a) nose position, pixels on the left edge of the picture; (b) key frames extracted by the ME-based method of Ref. 2; (c) key frames extracted by proposed method.



(a)

(b)

Fig. 4 Extracted key frames ("Foreman," CIF, key-frame ratio 2%): (a) ME-based method (frame numbers 0, 43, 107, 151, 171, 243, 299); (b) proposed method (frame numbers 0, 170, 178, 183, 195, 216, 299).

based method[2] and the proposed method are shown in Figs. 3(b) and 3(c), respectively. The first and the last frame were treated as default key frames in both methods. The frames extracted by the ME-based method[2] given in Fig. 3(b) are not the benchmark frames shown in Fig. 3(a) at all. This is because the frames that yield the local maximum ME when the head has the largest speed are with the head moving in a straight line, not at the inflection points of the track as in the key frames shown in Fig. 3(a). The frames extracted by the proposed method, as shown in Fig. 3(c), are almost the same as those in Fig. 3(a), which demonstrates the excellent subjective performance of the proposed acceleration-based method.

Figure 4 shows the key frames extracted from the "Foreman" sequence. Similarly to the results for "Erik," the key frames extracted by the proposed method are more distinct, and can describe the whole sequence better than those extracted by the ME-based method.

From the analysis, it can be concluded that the key frames selected by the proposed method are more consistent with the key frames determined by human perception than those selected by the ME-based method.

4 Conclusions

An acceleration-based key-frame extraction method is proposed by constructing a new factor that reflects the motion change of the primary moving object. Experimental results show that the proposed method selects key frames more consistent with human perceptions than the ME-based method.

Acknowledgment

We would like to acknowledge the support provided by the National Science Foundation of China (60772134) and the 111 Project (B08038).

References

1. B. T. Truong and S. Venkatesh, "Video abstraction: a systematic review and classification," in *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, archive 3(1), article 3 (2007).
2. T. Y. Liu, X. D. Zhang, J. Feng, and K. T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recogn. Lett.* **25**, 1451–1457 (2004).
3. W. S. Chau, O. C. Au, and T. S. Chong, "Key frame selection by macroblock type and motion vector analysis," in *2004 IEEE Int. Conf. on Multimedia and Expo*, Vol. **1**, pp. 575–578.
4. T. M. Liu, H. J. Zhang, and F. H. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Circuits Syst. Video Technol.* **13**(10), 1006–1013 (2003).