KEY FRAME EXTRACTION IN 3D VIDEO BY RATE-DISTORTION OPTIMIZATION

Jianfeng Xu, Toshihiko Yamasaki, and Kiyoharu Aizawa

Dept. of Electronics Engineering and Dept. of Frontier Informatics,
The University of Tokyo, Japan
{fenax, yamasaki, aizawa}@hal.k.u-tokyo.ac.jp

ABSTRACT

3D video, which consists of a sequence of 3D mesh models, can provide detailed 3D information both in spatial and temporal domain. In this paper, a key frame extraction method has been developed to summarize 3D video by rate-distortion optimization. For this purpose, we introduce an effective feature vector extraction algorithm from 3D video. Prior to key frame extraction, shot detection is performed using the feature vectors as a pre-processing. Then, a rate-distortion (R-D) curve is generated in each shot, where the locations of key frames are optimized. Lastly, R-D trade-off can be achieved by optimizing a cost function with a Lagrange multiplier. Our experimental results show the extracted key frames are compact and faithful to original 3D video.

1. INTRODUCTION

Research into three-dimensional (3D) imaging techniques has expanded dramatically in recent years because of its potential in many applications, such as in education, CAD, heritage documentation, broadcasting, and gaming. Recently, 3D video generation systems have been developed using multiple synchronous cameras [1, 2]. 3D video can record and reproduce faithful dynamic 3D information of real-world objects. It can reproduce not only shape and color of the real object but also temporal information (motion). Besides, 3D video is highly interactive since users can freely change the viewpoint. Fig. 1 shows some frames in a 3D video sequence generated by Tomiyama *et al.* [2], where each frame is expressed by a mesh model in VRML format.

The key frame extraction methods so far are for 2D video [3, 4] or motion capture data [5]. Key frame extraction is an efficient tool to summarize a video sequence [3, 4, 5, 6]. Many papers selected the key frames by measuring video content complexity [6]. In this paper, we optimize a cost function with both the rate (R) and distortion (D) to get R-D trade-off as the first attempt to key frame extraction from 3D video.

Similar to [3, 4], the prerequisite of our method is that shot detection should be done before key frame extraction, which is called shot-based method in [5]. Then, we define the rate and distortion in a simple but reasonable way so that we can get the R-D curves for each shot by optimizing the locations of key frames. Lastly, a cost function is derived by a Lagrange multiplier to get the R-D trade-off to determine the number of key frames. Since the algorithm is based on feature vectors, the computational cost is low.

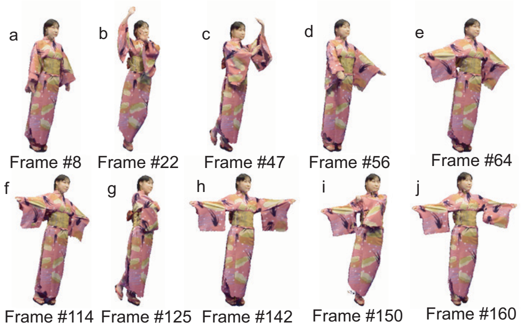

Fig. 1. Sample frames in a dancing 3D video sequence from a single view point.

2. SHOT DETECTION

We have developed two methods to segment 3D video into shots (some continuous frames in 3D video) by effective feature vectors [7, 8]. Both algorithms share the idea that 3D video is segmented by the motion (the amount of changes) of 3D object. A distance is calculated between two neighboring frames to reveal the motion after extracting the feature vectors for each frame. Then, 3D video is segmented by analyzing the distances with a decision strategy. In [7], three histograms of the distances of vertices from some fixed reference points are formed as the feature vectors. In [8], on the other hand, the feature vectors are based on three histograms of all the mesh vertices in spherical coordinate system, where the vertex positions are transformed from the Cartesian coordinate system to the spherical coordinate system. The shots are detected by analyzing the Euclidean distances of feature vectors as shown in Eq. (1).

$$d(m) = \sqrt{d^2\left(r,m\right) + d^2\left(\theta,m\right) + d^2\left(\phi,m\right)} \qquad (1)$$

where $d(m)$ denotes the distance between the m-th frame and

This work is supported by Ministry of Education, Culture, Sports, Science and Technology, Japan. All the 3D videos are provided by NHK, Japan.

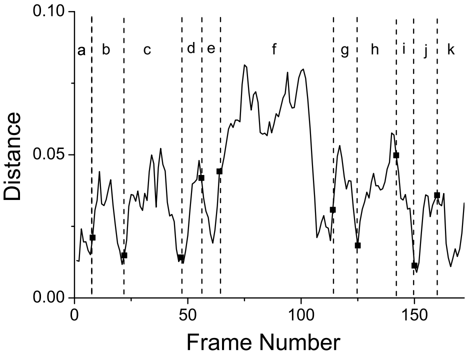

Fig. 2. Distance between two neighboring frames and detected shots in a dancing 3D video by [8].

the (m+1)-th frame, and $d(r, m)$, $d(\theta, m)$, $d(\phi, m)$ denote the Euclidean distances for three histograms between the m-th frame and the (m+1)-th frame in spherical coordinate system. Since r, θ, and ϕ reflect different types of information, namely, distance and angle information, it is required that similar motions in r, θ, and ϕ should cause similar distances in $d(r, m)$, $d(\theta, m)$, and $d(\phi, m)$. This requirement is satisfied by modifying the bin sizes of histograms [8].

In this paper, we utilize the distance in Eq. (1) and the ground truth of shot detection in [8]. The ground truth is based on eight independent assessors in the proposed evaluation approach [8]. Fig. 2 shows the distances and shots in a 3D video sequence used in our experiments and Fig. 1 shows the shot boundaries in Fig. 2.

3. KEY FRAME EXTRACTION

After shot detection, we will extract the key frames in each shot so that the whole key frames are composed by those in each shot. In this section, the rate and the distortion are defined. And R-D curves are generated by optimizing the key frame locations. Then, R-D trade-off can be achieved by a cost function with a Lagrange multiplier.

3.1. Definitions of Rate and Distortion

Generally, the rate should be the entropy of key frames and the distortion should be the information loss between the key frames and the original 3D video. In practice, we should give simple but reasonable definitions of the rate and distortion. Obviously, if we don't select any key frame, all the information will be lost, i.e., the distortion is the largest while the rate is the smallest. Conversely, if we use every frame as a key frame, all the information will be kept, i.e., the distortion is the smallest while the rate is the largest.

For simplification, the rate in a shot is defined as the number of key frames in a shot,

$$R(Shot_k) = I_k \tag{2}$$

where I_k denotes the number of key frames in the shot $Shot_k$, and $R(Shot_k)$ denotes the rate of the shot $Shot_k$.

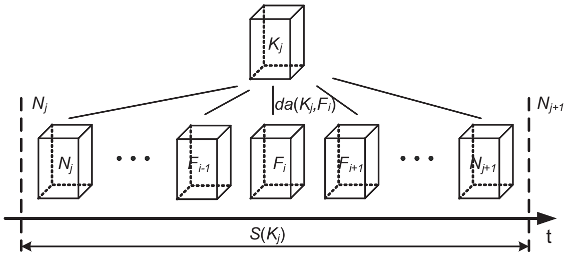

Fig. 3. Distortion definition for a key frame K_j.

In our definition, two aspects of the distortion are considered, i.e., the spatial distortion and the temporal distortion, respectively. The former comes from the spatial difference between a missed frame and its correspondent key frame, and the latter comes from the number of missed frames, or, how many frames a key frame represents in its interval. For example, if some continuous frames, which are delegated by only one key frame, are very different with each other, we will feel a lot of "distortion" which comes from the spatial distortion. And if too many frames are represented by a single key frame, we will feel "distortion" from the temporal distortion. Therefore, we define the distortion as Eq. (3), assuming that the distance can be summed.

$$Distortion(Shot_k) = \sum_{K_j \in Shot_k} Distortion(K_j) \tag{3}$$

$$Distortion(K_j) = \sum_{F_i \in S(K_j)} da(K_j, F_i) \tag{4}$$

$$da(K_j, F_i) = \sum_{m=\min(K_j,F_i)}^{\max(K_j,F_i)} d(m) \tag{5}$$

where $Distortion(K_j)$ denotes the distortion of the K_j-th key frame as shown in Fig. 3, $da(K_j, F_i)$ denotes the cumulative distance between K_j-th frame and F_i-th frame, which shows the spatial distortion. K_j is the j-th key frame, F_i is the i-th frame. $S(K_j)$ denotes the group of frames represented by the K_j-th key frame. $d(m)$ is the distance of feature vectors, which is shown in Eq. (1). $Shot_k$ is the k-th shot in 3D video. Equation (3) defines the total distortion for a shot $Shot_k$, where the sum of $da(K_j, F_i)$ is designed for the temporal distortion. If no key frame is selected in a shot, $Distortion(Shot_k)$ is the sum of all $da(K_j, F_i)$ in the shot.

3.2. R-D Curves

The main idea to achieve R-D curves is to find the minimum distortion for each given rate R by optimizing the locations of key frames. The optimal solution in 2D video was given by Lagendijk *et al.* [4], in which the boundaries of the intervals and location of the key frame within each interval were determined using an iterative algorithm (similar to the Lloyd-Max algorithm in scalar quantizer). We propose a sub-optimal strategy in this paper. Since 3D video is segmented by the mo-

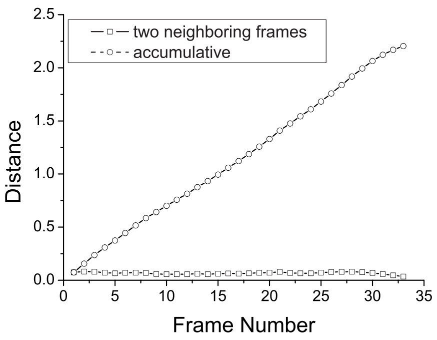

Fig. 4. Cumulative distances in a shot (shot “f” in Fig. 2).

tion of 3D object [8], the motion in a shot is similar. Therefore, suppose that $da(K_j, F_i)$ is linear, which holds well in our case as shown in Fig. 4. Then,

$$\begin{aligned} Distortion(K_j) &= \sum_{F_i=N_j}^{N_{j+1}-1} da(K_j, F_i) \\ &= \sum_{F_i=N_j}^{N_{j+1}-1} a|K_j - F_i| \qquad (6) \end{aligned}$$

$$N_1 = \inf Shot_k, \quad N_{R+1} = \sup Shot_k + 1 \qquad (7)$$

where a is a constant coefficient, N_j denotes the boundary of $S(K_j)$ as shown in Fig. 3. Therefore,

$$\begin{aligned} Distortion(K_j) = a[K_j^2 - (N_j + N_{j+1} - 1)K_j \\ +(N_j^2 - N_j + N_{j+1}^2 - N_{j+1})/2] \qquad (8) \end{aligned}$$

Assuming K_j is a continuous variable, partial differential should be calculated to get the minimum $Distortion(Shot_k)$. Since K_j is independent, we can get

$$K_j = \frac{N_j + N_{j+1} - 1}{2} \qquad (9)$$

This result is very intuitive, which is in the middle of the two boundaries as shown in Fig. 3. And the result has no relation with the distances in a shot but the linear assumption.

Similarly, we can get N_j

$$\begin{aligned} N_j &= \frac{K_{j-1} + K_j + 1}{2}, \quad j = 2, ..., R \qquad (10) \\ N_1 &= \inf Shot_k, \quad N_{R+1} = \sup Shot_k + 1 \qquad (11) \end{aligned}$$

Finally, the distortion can be calculated by Eq. (3). By calculating all the possible rates, R-D curves can be generated as shown in Fig. 5. It is observed our sub-optimal solution achieves no worse performance than the optimal iterative algorithm as shown in Fig. 6 since the optimal algorithm assumed the continuous variable in temporal domain, which does not exactly hold.

3.3. R-D trade-off

The main idea for R-D trade-off is that to summarize 3D video compactly requires the rate should be as small as possible. On the other hand, to summarize 3D video faithfully means the distortion should be as small as possible. However, these two requirements are conflicting as you can see in Fig. 5. Therefore, the R-D trade-off is necessary, which will be achieved according to the R-D curves for all the shots. A cost function with a Lagrange multiplier shown in Eq. (12) is optimized for each shot (Also shown in Fig. 5). The maximum rate is defined as 1/4 of the frame number in a shot. Also, at least one key frame will be selected in a shot. The users can decide Lagrange multiplier λ according to their wishes. We would like to mention that the R-D curve reveals the relationship between R and D so that the cost function is only determined by R. Fig. 7 shows some cost curves for different λ in a shot, where the black dots show the minimum cost values. From Fig. 7, we can see the rate will be smaller if λ is larger, that is to say, fewer key frames will be selected.

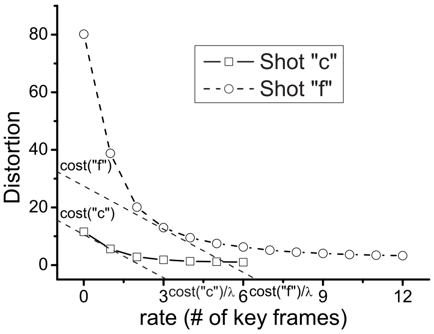

Fig. 5. Two R-D curves (shot “c” and shot “f” in Fig. 2).

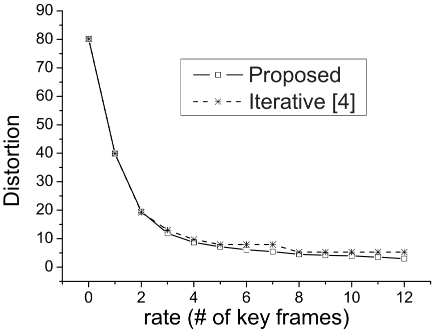

Fig. 6. Comparison in R-D curves (shot “f” in Fig. 2).

$$\begin{aligned} cost_\lambda(Shot_k) &= Distortion(Shot_k) + \lambda R(Shot_k) (12) \\ R^*(Shot_k) &= \arg \min_{R(Shot_k)} (cost_\lambda(Shot_k)) \qquad (13) \end{aligned}$$

Since Lagrange multiplier λ in all the shots is the same, the rates in those shots with fewer frames or smaller motions will be smaller as shown in Fig. 5. That is to say, this strategy selects fewer key frames in shorter shots or smaller motion shots, which is rather reasonable. Therefore, our method will automatically decide the key frame number in each shot by R-D optimization so that users can avoid to pre-decide the key frame number in each shot, which is a difficult task.

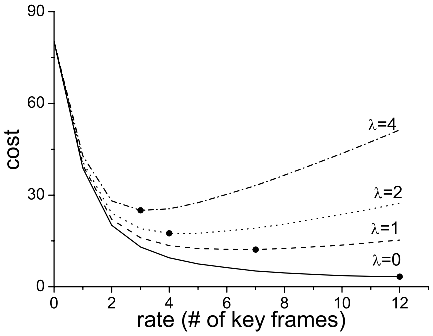

Fig. 7. Cost function with different λ in shot “f” of Fig. 2.

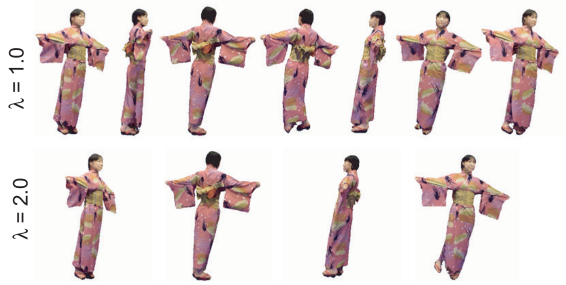

Fig. 8. Key frames in a rotation shot (shot “f” in Fig. 2). Top: $\lambda = 1.0$; Bottom: $\lambda = 2.0$.

4. EXPERIMENTAL RESULTS

Fig. 2 shows the distances from shot detection algorithm [8] in a typical 3D video sequence with 173 frames. Fig. 8 shows the key frames in a rotation shot (shot “f” in Fig. 2) by two Lagrange multipliers, which shows the influence of λ. As demonstrated in Fig. 7, fewer key frames will be selected if λ is larger. Fig. 9 shows the key frames in the whole sequence with $\lambda = 2.0$, whose distortion is 50.98 and rate is 15. From Fig. 9, our method can summarize 3D video well.

It is difficult to evaluate the experimental results. Our results come from the trade-off between the rate and the distortion as we have defined. However, the evaluation should be given by the users themselves. And there are no benchmarking or ground truth results for key frame extraction so far. Also, the best Lagrange multiplier λ should differ among different users, which is highly subjective. Anyway, the result evaluation may be an interesting topic for our future work.

5. CONCLUSIONS

In this paper, we have proposed a method to extract the key frames in 3D video. We set up an R-D model after defining the rate and the distortion in a simple and reasonable way. With two assumptions, an analytic solution is derived, which optimizes the distortion in a given rate and achieves no worse performance than the iterative algorithm. Then, a cost function is optimized, which considers the trade-off between the rate and the distortion with the Lagrange multiplier λ. In our method, the first optimization determines the key frame locations and the second optimization determines the key frame number, which is different from most reported algorithms. Experimental results show the key frames are compact and faithful to the original 3D video.

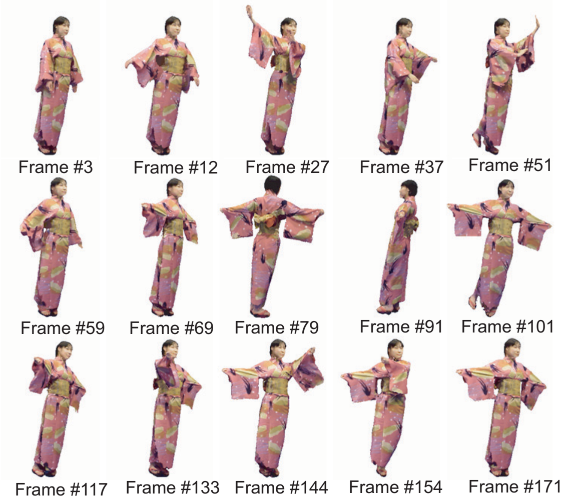

Fig. 9. All the key frames in a dancing 3D video with $\lambda = 2.0$.

6. REFERENCES

[1] T. Kanade, P. Rander, and P. Narayanan, “Virtualized reality: constructing virtual worlds from real scenes,” *IEEE Multimedia,* Vol. 4, No. 1, pp. 34–47, Jan./Mar. 1997.

[2] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadat., “Algorithm for dynamic 3D object generation from multiviewpoint images, ” *Proc. of SPIE,* Vol. 5599, pp. 153–161, 2004.

[3] H.S. Chang, S. Sull, and S.U. Lee, “Efficient Video Indexing Scheme for Content-Based Retrieval,” *IEEE Trans. CSVT,* Vol. 9, No. 8, pp. 1269–1279, Dec. 1999.

[4] R.L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, and E. Persoon, “Visual Search in a SMASH System,” in *Proc. IEEE ICIP’96,* Vol. 3, pp. 671–674, Sept. 1996.

[5] F. Liu, Y. Zhuang, F. Wu, and Y. Pan, “3D motion retrieval with motion index tree,” *Computer Vision and Image Understanding,* Vol. 92, No. 2/3, pp. 265–284, Nov./Dec. 2003.

[6] P. Aigrain, H. Zhang, and D. Petkovic, “Content-based representation and retrieval of visual media: A state-of-the-art review,” *Multimedia Tools Applicat.,* Vol. 3, pp. 179–202, 1996.

[7] J. Xu, T. Yamasaki, and K. Aizawa, “3D Video Segmentation Using Point Distance Histograms, ” in *Proc. IEEE ICIP’05,* pp. I-701–I-704, Italy, Sept. 2005.

[8] J. Xu, T. Yamasaki, and K. Aizawa, “Effective 3D Video Segmentation Based on Feature Vectors Using Spherical Coordinate System, ” *Meeting on Image Recognition and Understanding,* pp. 136–143, Awaji, Japan, Jul. 2005.