DEPTH PROPAGATION IN 2D-TO-3D CONVERSION BASED ON FRAME CLUSTERING

Zhenxiao Fu[2], Jiande Sun[1,2], Qiang Wu[2], Jing Li[1,3]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan China
[2] School of Information Science and Engineering, Shandong University, Jinan China
[3] School of Mechanical and Electrical Engineering, Shandong Management University, Jinan China

ABSTRACT

Depth propagation plays an important role to improve the quality of converted video in 2D-to-3D conversion. During depth propagation, the propagation path is usually along the time. In this paper, a new depth propagation algorithm is proposed to improve the quality of depth propagation by changing the propagation path according to frame clustering. The frames are clustered via K-means clustering based on HSV color histogram. The frame at the center of each cluster is selected as the keyframe of the cluster and the other frames are taken as the non-keyframes attaching to the keyframe. The depth information of keyframe is propagated to the non-keyframes within the same cluster directly. The performance of the proposed depth propagation algorithm is evaluated by MSE of the propagated depth map comparing with several existing algorithms. The comparison results show that the depth propagation errors can be reduced a lot by the proposed clustering based propagation.

Index Terms—2D-to-3D conversion, depth propagation, propagation path, keyframe selection, frame clustering

1. INTRODUCTION

In the last two decades, 2D-to-3D conversion was a hot spot in the field of 3D video. 2D-to-3D conversion algorithms are categorized into three types: manual, semi-automatic, and automatic 2D-to-3D conversion and how to improve the quality of the converted 3D video is one of the hot issues concerned widely in all of the three kinds of conversion. Manual 2D-to-3D conversion can obtain 3D video with high quality, but it costs extremely much labor power and time. On the other hand, though the automatic conversion needs no manual interaction, the quality of converted 3D video is usually unacceptable. Semi-automatic 2D-to-3D conversion attracted increasing attentions as it balanced the quality of converted 3D video and manual interaction [1].

Semi-automatic and automatic 2D-to-3D conversions have almost the same framework as shown in Fig. 1, which consists of: keyframe selection [2-4], depth assignment [5-6], depth propagation [7-8] and depth image based rendering (DIBR) [9]. In this framework, semi-automatic 2D-to-3D conversion is different from automatic conversion in keyframe selection and depth assignment on keyframes. Automatic conversion usually selects the keyframes at uniform temporal intervals along the video due to the computational complexity without considering the variation of scene [6] and estimates the depth based on depth cues [5]. Semi-automatic conversion needs manual interactions on these two parts. As the depth information is propagated from the keyframe to the non-keyframes via the methods of depth propagation, the keyframe selection and depth assignment with manual interactions are helpful to reduce the error of depth propagation, which can produce converted 3D video with high quality.

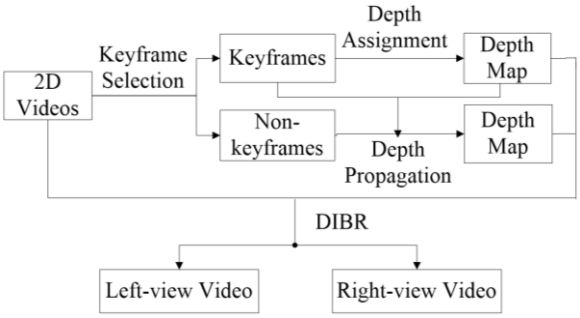

Fig.1 The framework of (semi-) automatic 2D-to-3D conversion

Recently, in order to keep the quality of 2D-to-3D conversion with as little manual labor as possible, there have been some exploring studies on automatic keyframe selection algorithms for 2D-to-3D conversion. Sun et al proposed to use the occlusion area between two consecutive frames to measure their difference in depth and select the keyframes according to the accumulated occlusion area [3]. Xie et al proposed to select keyframes from the view of depth estimation and selected the frames with distinct depth cue as keyframes [4]. Ju et al proposed to select keyframes according to color variation and object motion to reduce the depth propagation errors [10]. The depth propagation based on these algorithms reduced the errors of depth propagation and improved the quality of converted 3D video.

Though the above-mentioned algorithms select out the most representative frame for the depth structure of the

scene, the similarity in depth structure is not considered in the selection of non-keyframes. As the depth information of a frame should be propagated to the frame with most similar depth structure during depth propagation in order to keep the propagation errors as few as possible. The propagation error is related to the distance in depth structure between two propagated frames. That is to say, the propagation error will be high if the propagation distance is great, and vice versa. In existing depth propagation algorithms, the propagation distance is usually defined as the temporal distance between two frames. Though such definition works for most of the cases, it is not always the correct case. For instance, two consecutive frames may be quite different if they are in different scenes, the same scene repeats frequently, and two frames may be quite similar when they are in the repeated scenes even if they are temporally far away from each other.

In this paper, a depth propagation algorithm is proposed to find a depth propagation path, in which the propagation distance is defined as the similarity between two frames based on HSV color histogram. The center of each cluster is selected as the keyframe and the depth of keyframe is propagated to its corresponding non-keyframes directly. Our contributions in this paper are: 1) K-means clustering algorithm is used to determine the keyframes and the associated non-keyframes, to which the depth information of keyframes should be propagated. The global feature, HSV color histogram, is used for clustering. 2) The depth information of keyframe is propagated to the non-keyframes directly. The keyframe and its corresponding non-keyframes are in the same cluster and not limited in the same temporal segment. The experimental results indicate that the proposed algorithm is feasible and reliable for depth propagation.

2. FRAME CLUSTERING

Keyframe selection plays an important role in 2D-to-3D conversion as it can balance the depth propagation quality and manual interaction. In 2D-to-3D conversion, more keyframes usually bring higher depth propagation quality and require more manual interaction in depth assignment. The automatic keyframe selection attracted increasing attention in recent years, though keyframe selection on uniform temporal intervals is the mostly common used method. Some works studied to reach the same or better quality with less keyframes via keyframe selection [2-4].

Frame clustering is one of keyframe selection and video summary methods in video content analysis. HSV color histogram is a simple but effective global feature, which is usually used in frame clustering. In this paper, the 512 bin HSV color histogram for each frame is constructed, and a distance based on color difference is defined and used for frame clustering. K-means clustering is adopted as it is a kind of typical method of unsupervised learning [11] and it can simplify the clustering procedure without any artificial labelled sample.

2.1. HSV color histogram

In this paper, the three components in HSV color space, H, S and V, are non-uniformly quantized according to the color sensibility of human visual system (HVS). Each of the three components is quantized into eight parts as follows:

$$H=\begin{cases}0, H\in(316,20]\\ 1, H\in(20,40]\\ 2, H\in(40,75] \quad \begin{cases}S=n, S\in[0*n, 0.125*(n+1)]\\ V=n, V\in[0*n, 0.125*(n+1)]\end{cases}\\ 3, H\in(75,155] \\ 4, H\in(155,190] \quad n=\{0,1,2,3,4,5,6,7\}\\ 5, H\in(190,270]\\ 6, H\in(270,295]\\ 7, H\in(295,315]\end{cases} \tag{1}$$

These quantized components are transformed to the one dimensional feature vector, $L=64H+8S+V$, which can be used to obtain the 512 bin HSV color histogram.

2.2. K-means clustering

The number of clusters K is the same as the number of keyframes and K frames are randomly selected as the initial cluster centers from the video. The proposed distance between two frames used in K-means clustering is based on HSV color histogram and denoted as d. d is initialized as 0 and calculated as follows:

for $m=1:512$

$$d=\begin{cases}d, \left|h_{f_i}(m)-h_{f_j}(m)\right|\le T\\ d+1, \left|h_{f_i}(m)-h_{f_j}(m)\right|>T\end{cases} \tag{2}$$

end

where $h_{f_i}(m)$ denotes the mth bin of the color histogram of ith frame. T is the threshold of the bin difference.

The calculation of d shows that the proposed distance based on HSV color histogram concentrates on the color proportion difference between two frames. If the difference is larger than T, the two frame is considered different in the corresponding color. If there are too many color differences, the corresponding two frames are considered to be different. Fig. 2 shows the relationship between the proposed distance based on HSV color histogram and the corresponding MSE between groundtruth depth and the propagated depth on non-keyframes. The relationship can be fitted as shown in the black curve and there are 475 points in Fig. 2. It can be concluded that the proposed distance based on HSV color histogram can reflect the depth MSE between two frames.

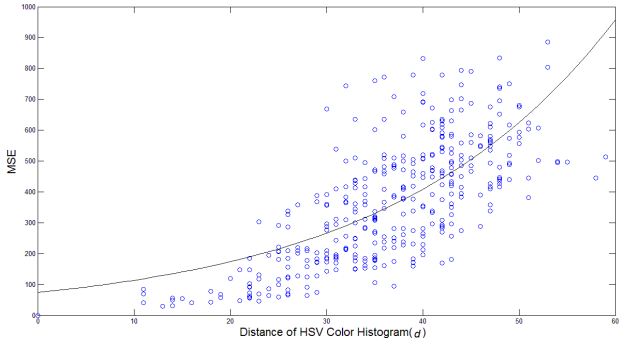

Fig. 2 The relationship between MSE of propagated depth map and the distance based on HSV color histogram

2.3. Depth propagation based on frame clustering

Fig. 3 shows an example result of video frame clustering and non-keyframe sorting in video "Breakdancer". X-axis denotes the clusters and y-axis denotes the temporal order of the frames. The red circle represents the keyframe in each cluster and the blue circles represent the non-keyframes in each cluster. It can be seen that the keyframe is not the first frame of each cluster, while the first frame of each cluster is usually selected as the keyframe conventionally. The non-keyframes in the same cluster are not temporally continuous.

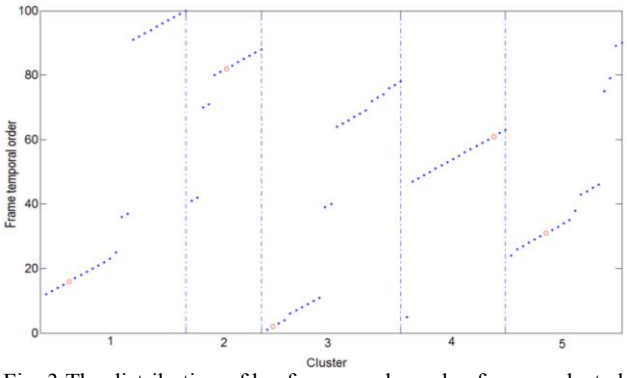

Fig. 3 The distribution of keyframes and non-keyframes selected by the proposed method for the video "Breakdancer".

After clustering, the center of each cluster is assigned as the keyframe and the other frames in the same cluster are taken as the non-keyframes attached to the keyframe. The depth of keyframes is propagated to its corresponding non-keyframes directly. Fig. 4 shows the proposed propagation algorithm in Fig. 4(c) comparing with the two most common used propagation methods via an example with one keyframe and three non-keyframes. The depth information of keyframe is propagated to each non-keyframe directly in Fig. 4(a) and propagated from keyframe to non-keyframe and non-keyframe to non-keyframe with time order in Fig. 4(b). The difference between Fig. 4(a), (b) and (c) is that there is no temporal limitation during depth propagation in Fig. 4(c).

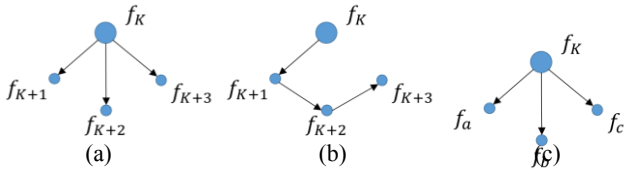

Fig. 4. Examples of different depth propagation paths.

3. EXPERIMENTAL SETTING

3.1. Experimental data

In order to verify the feasibility and reliability of the proposed depth propagation algorithm based on frame clustering (FC), three standard videos and part of American sitcom, The Big Bang Theory (TBBT), are used as the experimental videos. The four standard videos are "Kendo" with 300 frames, "Breakdancer" with 100 frames, and "Ballet" with 100 frames and the part of video "TBBT" is with 150 frames. The depth of video "Kendo" is the same produced in [3-4] and the other two standard videos have groundtruth depth. The depth of video "TBBT" is assigned manually and used as the groundtruth depth in the performance evaluation as there is no groundtruth depth for video "TBBT".

3.2. Depth propagation

The depth structures of the keyframes in different clusters are considered different from each other, so the depth information of each keyframe is propagated to the non-keyframes in the same cluster unidirectionally. The shifted bilateral filtering (SBF) method proposed in [6] is adopted as it is the most common used one in 2D-to-3D conversion.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed propagation algorithm based on keyframe clustering (KC), four keyframe selection algorithms are adopted in the experiments, which are the algorithms based on uniform temporal interval (UTI) [6], accumulated histogram (AH) [12], accumulated occlusion (AO) [3] and depth estimation (DE) [4]. Each algorithm selects the same number of keyframes, which is set according to the ratio of every 25 frames one keyframe. However, as the first frame of each segment is always selected as a keyframe in the algorithms used for comparison, if the video has 100 frames, there should be 5 keyframes. The comparison results on average MSE in the experimental videos are listed listed in TABLE I-TABLE II. In TABLE I, the column of MSE1 lists the average MSE comparison between the proposed algorithm and the other algorithms with propagation shown as Fig. 4(a). The column of MSE2 lists the average MSE comparison between the proposed algorithm and the other algorithms

with propagation shown as Fig. 4(b). It can be seen that in MSE1 comparison, the improvement of the proposed algorithm over the best of the other algorithms in propagation error is from **13.5%** to **40.6%**, and in MSE2 comparison, the corresponding improvement is from **5.9%** to **15.5%**. It can be seen that the proposed FC algorithm can reach the lowest MSE than the others.

TABLE I COMPARISON ON AVERAGE MSE BETWEEN ALGORITHMS WITH THE SAME NUMBER OF KEYFRAMES

Video	Method	Keyframes	MSE1	MSE2
Kendo	FC	12 31 42 58 80 105 119 135 156 181 208 243 280	**468.9**	
	UTI	1 25 50 75 100 125 150 175 200 225 250 275 300	**788.7 (40.6%)**	525.3
	DE	1 22 42 79 96 115 136 162 189 214 240 267 293	816.6	**520.3 (10.7%)**
	AH	1 27 42 99 115 151 170 208 224 250 261 274 298	969.1	599.2
	AO	1 24 45 70 100 120 145 169 193 220 243 264 290	874.5	568.0
Breakdancer	FC	2 16 31 61 82	**298.1**	
	UTI	1 25 50 75 100	391.2	341.9
	DE	1 26 49 74 96	368.5	322.7
	AH	1 15 50 69 89	384.5	401.5
	AO	1 27 48 75 95	**349.6 (14.7%)**	**316.8 (5.9%)**
Ballet	FC	12 37 45 57 86	**126.8**	
	UTI	1 25 50 75 100	195.6	151.0
	DE	1 26 50 73 95	182.2	144.2
	AH	1 12 42 57 93	**160.2 (20.8%)**	157.9
	AO	1 20 49 70 94	187.5	**150.0 (15.5%)**
TBBT	FC	3 16 59 67 85 103 132	**1482.4**	
	UTI	1 25 50 75 100 125 150	6861.7	4996.4
	AH	1 6 17 65 80 81 112	**1714.6 (13.5%)**	**1616.1 (8.27%)**

TABLE II lists the comparison between FC and UTI on the number of keyframes with comparable average MSE. It can be seen that the proposed FC algorithm can save from **16.7%** to **78.1%** keyframes over UTI on different videos.

5. CONCLUSIONS AND ANALYSIS

In this paper, the depth propagation algorithm based on frame clustering is studied, in which HSV color histogram is used for frame clustering to determine the keyframes and the depth information of the keyframes is propagated to the non-keyframes in the same cluster without considering the temporal order. The experimental results prove that frame clustering is helpful to reduce the propagation errors, which can obtain the converted 3D video with higher quality.

In addition, it can be concluded from the comparison in TABLE I, the propagation method shown in Fig. 4(b) can reach less propagation errors than that in Fig. 4(a). It means that the sorting of non-keyframes is useful for the improvement of the quality of depth propagation. So it will be the next point in our future study.

TABLE II COMPARISON BETWEEN FC AND UTI ON THE NUMBER OF KEYFRAMES WITH CORRESPONDING MSE

	NKS		UTI	
	Number of Keyframes	MSE	Number of Keyframes	MSE
Kendo	13	468.9	16***(18.8%)***	481.3
Breakdancer	5	298.1	6***(16.7%)***	293.3
Ballet	5	126.8	7***(28.6%)***	130.3
TBBT	7	1482.4	32***(78.1%)***	1483.4

6. ACKNOWLEDGEMENT

The work is supported by DAAD, the National Science Foundation of Shandong Province (ZR2014FM012), the National Natural Science Foundation of China (61305060, 61571274), Jinan Youth Star of Science and Technology Plan (201406002), and the Young Scholars Program of Shandong University (YSPSDU) (2015WLJH39). The contact author is Jiande Sun (e-mail: jiandesun@hotmail.com).

7. REFERENCES

[1] L. Zhang, C. Vazquez and S. Knorr, "3D-TV Content Creation: Automatic 2D-to-3D Video Conversion", IEEE Trans. on Broadcasting, 57(2), 2011, 372-383.

[2] Dichangsheng Wang, Ju Liu, Jiande Sun, et al, "A Novel Key-Frame Extraction Method for Semi-Automatic 2D-to-3D Video Conversion", Proceedings of 7th IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2012, 1-5.

[3] J. Sun, J. Xie, J. Li, et al, "A Key-Frame Selection Method for Semi-automatic 2D-to-3D Conversion", Proceedings of the 9th International Forum on Digital TV and Wireless Multimedia Communication (IFTC), 2012, 465-470.

[4] J. Xie, J. Sun, J. Liu, et al, "Key-Frame Selection Strategy Based on Edge Points Classification in 2D-to-3D Conversion", Proceedings of International Conference on Intelligence Science and Big Data Engineering (IScIDE), LNCS 8261, 2013, 797-804.

[5] A. Torralba and A. Oliva, “Depth Estimation from Image Structure”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(9), 2002, 1226-1238.
[6] X. Cao, Z. Li and Q. Dai, “Semi-Automatic 2D-to-3D Conversion Using Disparity Propagation”, IEEE Trans. on Broadcasting, 57(2), 2011, 491-499.
[7] V.H. Philip, F. Julie, F. Simon, et al, “Rapid 2D-to-3D Conversion”, Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems IX, 2002.
[8] E. Tolstaya, P. Pohl and M. Rychagov, “Depth Propagation for Semi-Automatic 2D to 3D Conversion”, Proceedings of SPIE, Three-Dimensional Image Processing, Measurement (3DIPM), and Applications, 2015.
[9] F. Christoph, “Depth-Image-Based Rendering (DIBR), Compression, and Transmission for a New Approach on 3D-TV”, Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems XI, 2004.
[10] K. Ju and H. Xiong, “A Semi-Automatic 2D-to-3D Video Conversion with Adaptive Key-Frame Selection", Proceedings of SPIE, Optoelectronic Imaging and Multimedia Technology III, 2014.
[11] M. James, “Some Methods for Classification and Analysis of Multivariate Observations”, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(14), 281-297.
[12] B. Gunsel, A. Tekalp and P. Van Beek, “Moving Visual Representations of Video Objects for Content-Based Search and Browsing”, Proceedings of the 4th International Conference on Image Processing (ICIP), 2, 1997, 502-505.