I-039

An Evaluation Approach for Temporal Segmentation of 3D Videos

Jianfeng Xu† Toshihiko Yamasaki‡ and Kiyoharu Aizawa‡

1. Introduction

Recently, 3D videos are attracting increasing attention by many researchers since 3D videos can provide much more realistic observation from any view point and be potentially applied in many applications. One significant difference from 3D animations, which are generated by computer, is that 3D videos are the recordings of the real world, captured by multi-cameras frame by frame [1, 2, 3]. To build a 3D video database [1, 4, 5], it is essential to segment a 3D video into shots for efficiently browsing, indexing, and retrieval. Similar to 2D video segmentation [6], 3D video segmentation is to divide 3D video in temporal domain into a set of meaningful and manageable shots that are used as basic elements for indexing. Here, a shot is several successive frames in a 3D video with some common visual characteristics, whose frames usually have similar motions in our task (shown in Figure 1). Although 3D videos have more dimensions than other media data, our previous works [4, 5, 9] show that it is possible to segment 3D video effectively and efficiently.

An important but still unsolved problem is how to evaluate the algorithm of 3D video segmentation. Performance evaluation of information retrieval system is difficult because the effectiveness is highly subjective and even depends on examinees and many other factors [10]. It is such a psychological problem that the straightforward way is just subjective evaluation by the examinees, which, however, is very expensive and time-consuming. Therefore, an evaluation approach for information retrieval systems is necessary [10].

Although the conventional evaluation measures, namely, *precision*, *recall*, and *F-measure* (briefly (P, R, F) in this paper), are adopted in our previous works [5, 9], it is necessary to pre-define a threshold (detailed in section 4.1), which directly affects (P, R, F). In this paper, a new evaluation approach for 3D video segmentation is proposed, which may have no threshold. To get the ground truth, several volunteers (or examinees) are asked to segment all the test sequences independently and then voting strategy is adopted. Our new evaluation approach is in a different view from the conventional one but covers the conventional one. Its efficiency is demonstrated by our experiments. Since the evaluation approach is quite general, it is also possible to apply it in other similar systems.

2. Related Work and Key Problems

Research on information retrieval has a long history [11]. At the same time, evaluation approaches have been also discussed carefully [12]. For evaluation measures, precision rate and recall rate are popular in information retrieval system, whose definitions are given in eqs. (1) and (2). *Recall* shows the ability to present all relevant items, and *precision* shows the ability to present only relevant items. Van Rijsbergen gave a good survey in Ref. [10], in which *E-measure* and *F-measure* are also defined as eqs. (3) and (4). Later, these measures are also widely applied in image retrieval [13], 2D video shot segmentation [14], and our previous works on 3D video segmentation [5, 9].

$$Recall = \frac{relevant\ correctly\ retrieved}{all\ relevant} \quad (1)$$

$$Precision = \frac{relevant\ correctly\ retrieved}{all\ retrieved} \quad (2)$$

$$E\text{-}measure = 1 - \frac{1}{\alpha\left(\frac{1}{Precision}\right) + (1-\alpha)\left(\frac{1}{Recall}\right)}, \alpha \in [0,1] \quad (3)$$

$$F\text{-}measure = \frac{1}{\alpha\left(\frac{1}{Precision}\right) + (1-\alpha)\left(\frac{1}{Recall}\right)}, \alpha \in [0,1] \quad (4)$$

Many papers on 2D video segmentation focus on detecting cuts and transitions for the special effects like fade-in, fade-out, and dissolve, which can be precisely detected by human beings so that the ground truth is easily obtained. However, in 3D video segmentation system, the shot boundary is rather subjective and blurry even for human beings according to our experiments, which is a key problem in our system. That is because our 3D video shot is divided by the motion of 3D object [5, 9] as shown in Figure 1. Figure 2 shows the shot boundaries of a pitching sequence by 8 volunteers, whose detail procedures are shown in section 3. From Figure 2, both the shot number and boundary sites are quite different among the volunteers. This infers it is not trivial to get a ground truth in 3D video segmentation.

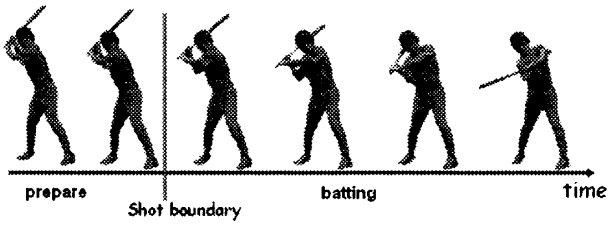

Fig. 1. One example for 3D video segmentation.

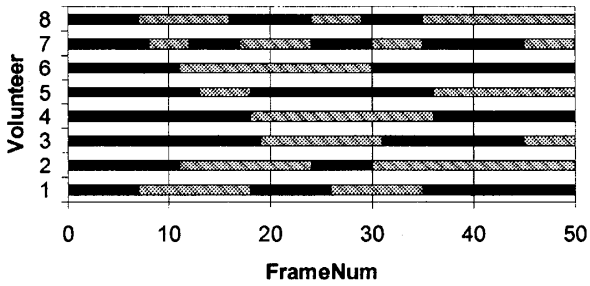

Fig. 2. Segmentation results for "Pitcher" by 8 volunteers.

† Dept. of Electronics Engineering, The University of Tokyo
‡ Dept. of Frontier Informatics, The University of Tokyo

3. Defining Ground Truth

For 3D video segmentation, we specially define (P, R, F) as:

$$R = \frac{relevant\ correctly\ retrieved}{all\ relevant} = \frac{n_{correct\_seg}}{n_{relevant\_seg}} \quad (5)$$

$$P = \frac{relevant\ correctly\ retrieved}{all\ retrieved} = \frac{n_{correct\_seg}}{n_{retrieved\_seg}} \quad (6)$$

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right)+(1-\alpha)\left(\frac{1}{R}\right)}, \alpha \in [0,1] \quad (7)$$

where R is *recall*, P is *precision*, and F is *F-measure*. $n_{correct\_seg}$ is the number of correct segmentation, $n_{relevant\_seg}$ is the number of relevant segmentation, which is also the ground truth, and $n_{retrieved\_seg}$ is the number of retrieved segmentation by our algorithm. α is a constant and is set as 0.5.

The ground truth $n_{relevant\_seg}$ comes from humans as other information retrieval systems do. As mentioned in section 2, since the relevant segmentations are highly subjective, more than one volunteer is asked to segment all the test data and the volunteers do not know others' segmentation results in advance. They will decide the segmentation criteria by themselves so that the results are rather high-level, which causes $n_{relevant\_seg}$ and shot boundary sites are independent of the others'. The volunteers use our own 3D video player to view the 3D video, whose view points are changeable. And each volunteer is provided printed video frames from frontal view (an example is shown in Figure 3), where the frame index is shown below its image. Each volunteer is asked to describe briefly the shots in his/her report. Totally, 3 sequences are segmented by 8 volunteers including "Toshiko" (a dancing sequence with 173 frames), "Pitcher" (a pitching sequence with 51 frames), and "Batter" (a batting sequence with 51 frames). All the test sequences are 10 frames per second. Figure 2 shows the segmentation results for "Pitcher" by 8 volunteers. To get the ground truth, we suppose two assumptions: one is the volunteers will keep their criteria for segmentation in the whole 3D video sequence; the other is there is some common ground for segmenting 3D video among all the volunteers. Therefore, the authors merge some neighboring shot boundaries from different volunteers and emerge one integrated result by considering of both the shot boundary sites and the volunteers' description on the shots. Then, the integrated result has two kinds of information as shown in Figure 4: one shows how to segment the 3D video and the other gives how many volunteers regard it as a shot boundary, which is called vote number in this paper. Therefore, $n_{relevant\_seg}$=$n_{relevant\_seg}(k)$ means the number of all relevant segmentation whose vote number is no less than k (or these segmentations are adopted by at least k volunteers).

4. Evaluation Approach

In this section, we firstly introduce the naive evaluation measures (P, R, F). Then, after discussing the drawbacks of (P, R, F), a new evaluation approach is presented in detail.

4.1 Naive Evaluation Measures

By eqs. (5) and (6), we need to calculate $n_{correct\_seg}$ for (P, R, F). $n_{correct\_seg}$ will amount those segmentations satisfied

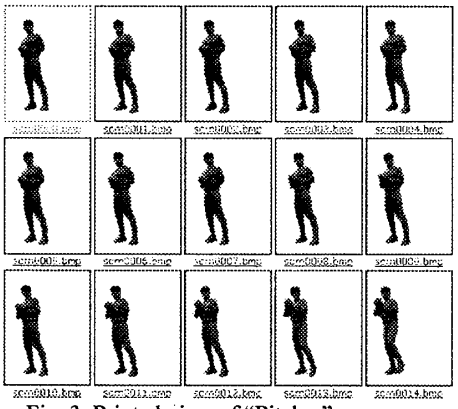

Fig. 3. Printed view of "Pitcher" sequence.

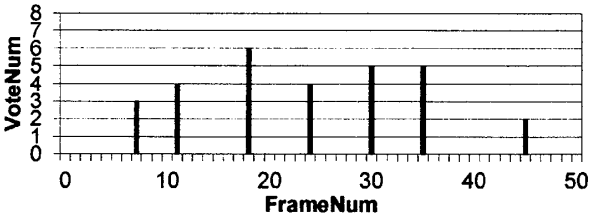

Fig. 4. Cluster results of segmentation of "Pitcher".

$$l_{relevant\_seg}(j) - \beta \le l_{retrieved\_seg}(i) \le l_{relevant\_seg}(j) + \beta \quad (8)$$

where $l_{relevant\_seg}(j)$ is the location (frame index) for the j-th relevant segmentation in $n_{relevant\_seg}(k)$, $l_{retrieved\_seg}(i)$ is the location for the i-th retrieved segmentation in $n_{retrieved\_seg}$, and β is a range considering the correctly retrieved. If β is 5, then it means the frames in 0.5 seconds will be considered as the same segmentation boundary since the test sequences are in 10 frames per second. Although β should depend on the sequences and the volunteers, 3 or 4 can be accepted in most of cases.

Then, *precision* and *recall* will be calculated as the function of vote number by eqs. (5) and (6). Generally, small vote number means more segmentation in ground truth so that *recall* will be small and *precision* will be large. In our experiment, vote number is set as 4 (that is to say, at least 50% volunteers).

4.2 New Approach Based on Set Theory

The evaluation approach above has some drawbacks in 3D video segmentation: one is that it is necessary to pre-define a threshold β to calculate (P, R, F); the other is only frame index difference is considered, which does not always match with the difference in perceptual quality. Figure 5 shows an example, where Pitcher #11 and Toshiko #56 are the shot boundaries by more than 50% volunteers, and Pitcher #6 and Toshiko #59 are the shot boundaries by one of our algorithms [5]. According to eq. (8), Pitcher #6 is not considered as correct segmentation while Toshiko #59 is considered as correct one although the two frames are more similar in the former case than in the latter case. When the motion of 3D object is great, the difference of two frames is rather large even if they are very near by frame index. It suggests that the same β has different meanings in different motion shots. In this section, new evaluation measures from set theory are proposed to avoid the drawbacks of conventional evaluation measures.

Pitcher #6 Pitcher #11 Toshiko #56 Toshiko #59

Fig. 5. Two examples for segmentation results.

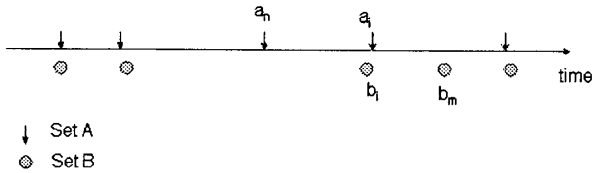

Fig. 6. Set A (ground truth) and Set B (from any algorithm).

We regard the segmentation results (shot boundaries) as frame set. Let A be the segmentation result from volunteers. Therefore, A is the ground truth, which can be obtained as described in section 3. Let B be the segmentation results of our segmentation method (or any algorithm of 3D video segmentation). Then, we need to define a "reasonable" "similarity measure" between set A and set B to evaluate the segmentation method (see Figure 6).

$$A = \{a_i, i = 0, ..., I-1\}$$
$$B = \{b_j, j = 0, ..., J-1\} \quad (9)$$

where a_i means that the a_i-th frame is the i-th segmentation regarded by volunteers, b_j means that the b_j-th frame is the j-th segmentation by our method.

Firstly, we define a measure to evaluate the over segmentation. The main requirement is the over segmentation will cause the measure larger while the missed segmentation will not affect it. Therefore, we define the difference between b_j and A as:

$$d(b_j, A) = \min_i(d(b_j, a_i)) \quad (10)$$

where $d(b_j, a_i)$ is a distance between the b_j-th and a_i-th frame. The minimum value will try to find the corresponding boundaries between A and B, and show their distance. Then, we average the difference $d(b_j, A)$ as the evaluation measure of over segmentation, which averages the distance of each boundary in set B.

$$d_{over} = \frac{1}{J}\sum_{j=0}^{J-1} d(b_j, A) = \frac{1}{J}\sum_{j=0}^{J-1} \min_i(d(b_j, a_i)) \quad (11)$$

From eq. (11), if there is an over segmentation b_m, $d(b_m, A)$ will be larger, refer to Figure 6. Therefore, d_{over} will become larger. The more over segmentations there are, the larger d_{over} would be. Then, d_{over} can reflect the over segmentation. Since the missed segmentation is located in set A, it will have no influence on d_{over}. This definition is corresponding to *precision*, which also evaluates the over segmentation.

Then, corresponding to *recall*, another measure is designed to evaluate the missed segmentation, which, similarly, is defined as the average difference of a_i and B.

$$d_{missed} = \frac{1}{I}\sum_{i=0}^{I-1} d(a_i, B) = \frac{1}{I}\sum_{i=0}^{I-1} \min_j(d(a_i, b_j)) \quad (12)$$

From eq. (12), if there is a missed segmentation a_n, $d(a_n, B)$ will be larger. Therefore, d_{missed} will become larger. The more missed segmentations there are, the larger d_{missed} will be. d_{missed} can reflect the missed segmentation. Since the over segmentation is located in set B, it will have no influence on d_{missed}.

Corresponding to *F-measure*, we define "*suitability*" as:

$$s = w_1 d_{over} + (1 - w_1) d_{missed} \quad (13)$$

where w_1 is a weighting, s is *suitability*. Note that different to (P, R, F), smaller the new measures are, better the performance is.

4.3 Analysis and Experiments

In this section, some analysis of the new evaluation is given, and the experiments on demonstrating the efficiency of the new evaluation are reported.

One interesting and important thing in the new evaluation is the definition of $d(a_i, b_j)$. A careful design of $d(a_i, b_j)$ can not only avoid the threshold problem but also achieve some other advantages. Two special cases are discussed: one can deduct the new evaluation to conventional one; the other can avoid the drawbacks of conventional one, which considers both the spatial distance and temporal distance.

In the first case, $d(a_i, b_j)$ is defined as:

$$d(a_i, b_j) = \begin{cases} 1 & |a_i - b_j| > \beta \\ 0 & |a_i - b_j| \le \beta \end{cases} \quad (14)$$

Then, according to eq. (8)

$$d_{over} = \frac{n_{over\_seg}}{J} = 1 - \frac{n_{correct\_seg}}{n_{retrieved\_seg}} = 1 - precision$$
$$d_{missed} = \frac{n_{missed\_seg}}{I} = 1 - \frac{n_{correct\_seg}}{n_{relevant\_seg}} = 1 - recall \quad (15)$$

where n_{over_seg} is the number of over segmentation, and n_{missed_seg} is the number of the missed segmentation. Therefore, *precision* and *recall* are special cases in our new evaluation measures.

In the second case, we consider frame contents by Euclidean distance. However, if two frames in set B are both near one frame in set A, which means an over segmentation in set B, this over segmentation is difficult to reflect by *only* Euclidean distance. To avoid this, $d(a_i, b_j)$ is defined as:

$$d(a_i, b_j) = |a_i - b_j| \times \left\| FV(a_i) - FV(b_j) \right\|_2 \quad (16)$$

where $FV(a_i)$ and $FV(b_j)$ are the feature vectors of the a_i-th frame and the b_j-th frame. In eq. (16), the first part is the frame index difference, which means the temporal distance, and the second part is the Euclidean distance, which means the spatial distance. This definition also avoids the mismatch between two frames in set A and set B, whose feature vectors are similar but frame indexes are very different. Such a mismatch will reduce the result.

It is found that there are strong correlations between the two evaluation approaches as shown in Figure 7. The average correlation between the two is 0.9335, detailed in Table 1. In this experiment, we randomly select 9 segmentation results in different parameters by our method [5]. The vote number is set as 4 in our experiment.

Briefly, the advantages of our evaluation measures are:

(1) We do not need to define the threshold β, which is necessary in *precision* and *recall*.

(2) *Precision* and *recall* are special cases in our new measures.

(3) The new evaluation is more accurate and reasonable.

Table 1 Correlations between two evaluation measures.

Evaluation Measure	Correlation
precision & d_{over}	0.9274
recall & d_{missed}	0.8967
F-measure & *Suitability*	0.9763
Average	0.9335

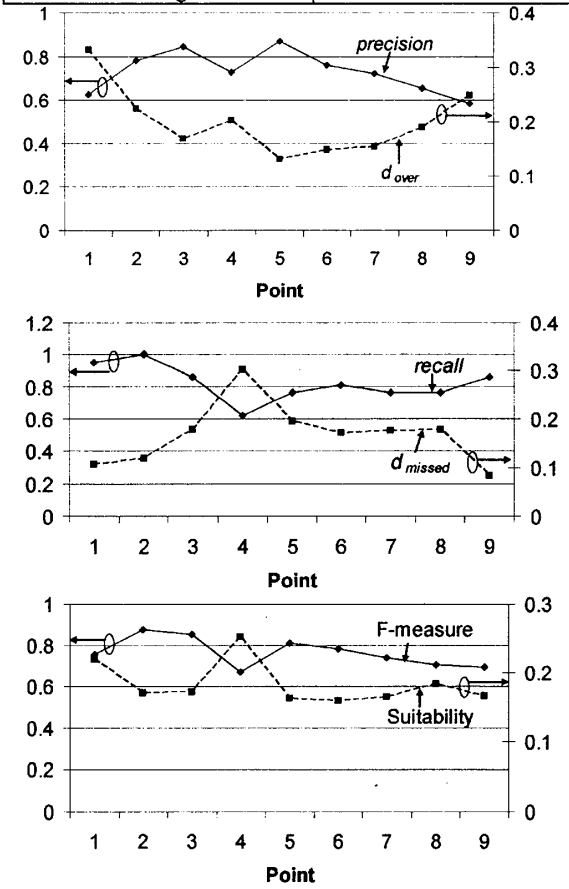

Fig. 7. Correlations between two evaluation measures.

We consider both the content of two frames and the frame index difference instead of only the frame index difference. In other words, both spatial distance (by Euclidean distance of *FV*s) and temporal distance (by frame index difference) are included in our evaluation measures, which is more complete.

5. Conclusion

In this paper, a new evaluation approach for 3D video segmentation has been proposed. We described in detail how to obtain the ground truth for 3D video segmentation system after pointing out the difficulty. Then, new evaluation measures based on set theory have been presented, whose efficiency was demonstrated by our experiments. Moreover, *precision* and *recall* are included as special cases of the proposed approach. The advantages of new evaluation approach have been discussed due to adopting both the spatial and temporal distances. Also, it is possible to apply them in other systems similar to 3D video segmentation without or with little modification.

6. Acknowledgements

All test data including "Toshiko", "Batter", and "Pitcher" are provided by NHK, which is greatly appreciated by the authors. This work is supported by Ministry of Education, Culture, Sports, Science and Technology under the "Development of fundamental software technologies for digital archives" project.

7. Reference

[1] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, "Algorithm for dynamic 3D object generation from multi-viewpoint images," Proceeding of SPIE, Vol. 5599, pp. 153-161, 2004.

[2] T. Matsuyama, X. Wu, T. Takai, and T. Wada, "Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video," IEEE Trans. Circuit and System for Video Technology, Vol. 14, No. 3, pp.357-369, March 2004.

[3] 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./March 1997.

[4] J. Xu, T. Yamasaki, and K. Aizawa, "3D Video Segmentation Using Point Distance Histograms," IEEE International Conference on Image Processing (ICIP), Genoa, Italy, Sept. 11-14, 2005 (accepted).

[5] 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 (MIRU) 2005, Awaji, Japan, July 18-20, 2005 (accepted).

[6] I. Koprinska and S. Carrato, "Temporal video segmentation: A Survey," Signal Processing: Image Communication, Vol. 16, No. 5, pp. 477-500, Jan. 2001.

[7] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," Proceedings of Shape Modeling Applications, 2004. pp: 145-156, 7-9 June 2004.

[8] Y. Aslandogan and C. Yu, "Techniques and Systems for Image and Video Retrieval," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, pp. 56-63, Jan./Feb. 1999.

[9] J. Xu, T. Yamasaki, and K. Aizawa, "Segmentation for 3D Video Archives," the 1st Digital Contents Symposium (DCS), Tokyo, Japan, May 25-27, 2005.

[10] C. J. van Rijsbergen, "Information retrieval," London: Butterworths, 1979.

[11] J.A. Swets, "Information retrieval systems," Science, Vol. 141, pp. 245-250, 1963.

[12] W. Goffman and V.A. Newill, "A methodology for test and evaluation of information retrieval systems," Information Storage and Retrieval, Vol. 3, pp. 19-25, 1966.

[13] R. Krishnapuram, S. Medasani, S.H. Jung, and et al, "Content-based image retrieval based on a fuzzy approach," IEEE Trans. on Knowledge and Data Engineering, Vol. 16, Issue 10, pp.1185-1199, Oct. 2004.

[14] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," Proceedings of SPIE, Vol 3656, pp. 290-301, 1999.