AN OPTIMAL FRAMEWORK FOR SUMMARIZATION OF STEREOSCOPIC VIDEO SEQUENCES

Nikolaos D. Doulamis, Anastasios D. Doulamis, Yannis S. Avrithis,
Klimis S. Ntalianis and Stefanos D. Kollias

Electrical and Computer Engineering Department,
National Technical University of Athens
E-mail: ndoulam@image.ntua.gr

ABSTRACT

In this paper an optimal framework for summarization of stereoscopic video sequences is presented, which extracts a meaningful set of video frames. Each sequence is first partitioned into shots, the disparity field, occluded areas and depth map are estimated and then a hierarchical color and depth segmentation scheme is applied to each shot, based on a multiresolution implementation of the RSST algorithm. Color and depth segment fusion is employed for achieving high-quality semantic segmentation, and feature vectors are constructed using a fuzzy classification formulation. For a given shot, key frames are extracted using an optimization method, namely, a genetic algorithm, for locating frames of minimally correlated feature vectors. Experimental results indicate the reliable performance of the proposed scheme.

1. Introduction

The use of three-dimensional (3-D) video has recently increased since it provides more efficient visual representation and enhances multimedia communication [1]. Three-dimensional video description enables users to handle and manipulate video objects more efficiently by exploiting, for example, the depth information, provided by stereo vision [2]. However, traditionally, 3-D (stereo or multiview) image sequences are recorded sequentially using slightly different viewpoints of the same scene. Such video representation has a number of limitations for the new emerging multimedia applications, such as video browsing, content-based indexing and retrieval. Currently, the only way to browse a video is to sequentially scan video frames, a process that is both time-consuming and tedious. Furthermore, video queries are insufficiently performed on entire video sequences, due to significant temporal redundancy of video content. So new methods for video content representation should be implemented [3].

Recently, some approaches have been proposed. In particular selection of a single key frame for each shot has been presented in [3], [4], which cannot provide sufficient information about the video content. Construction of compact image maps has been described in [5]. Additionally a method for analyzing video and building a pictorial summary for visual representation has been proposed in [6]. Although such approaches can be very efficient for specific applications they cannot provide very satisfactory results in real world complex shots. Moreover, all the aforementioned works are dealing with 2-D video sequences and cannot be directly applied to 3-D video archives since they do not exploit 3-D information.

In the context of this paper a generalized framework for non-linear representation of 3-D video sequences is proposed, regardless of the scene complexity. We accomplished this by the following steps: we merge color segments that belong to similar depth. Color segments give very accurate contours of the objects while segments of video objects are usually located on the same depth plane. To accelerate the color and depth segmentation process a multiresolution implementation of the Recursive Shortest Spanning Tree (RSST) algorithm is presented. All features extracted by the video sequence analysis module are gathered together using a fuzzy feature vector formulation to increase the robustness of the proposed summarization scheme. Finally, key frames within each shot are extracted by minimizing a cross correlation criterion by means of a genetic algorithm.

2. Analysis of Stereo Video Sequences

The analysis of the 3-D sequences starts by calculating the disparity field. Let us assume that the variable Z expresses the depth. Let us also consider as (x_1,y_1) and (x_2,y_2) two image points generated by the perspective projection of a 3-D point $\mathbf{w}$ onto the two image planes I_1 and I_2. In particular by denoting as $\boldsymbol{d}(x_1,y_1)$ the disparity vector at location (x_1,y_1) in camera 1 with respect to camera 2, the vector $\boldsymbol{d}(x_1,y_1) = [d_x(x_1,y_1)\ d_y(x_1,y_1)]^T$ is given by

$$d_x=d_x(x_1,y_1)=x_2-x_1= f_1(Z) \tag{1a}$$

$$d_y=d_y(x_1,y_1)=y_2-y_1= f_2(Z) \tag{1b}$$

Therefore if the disparity vector is known, (1) reduces to an overdetermined linear system of two equations with a single unknown, i.e, *Z*, and a least-squares solution can be obtained.

Although computation of depth from disparity is straightforward, the estimation of disparity field from the images on planes I_1 and I_2, is an elaborate task that involves matching of each point (x_1,y_1) on I_1 with a corresponding point (x_2,y_2) on I_2, resulting in a high computational cost. Disparity estimation is accomplished by means of a block matching algorithm, similar to the one described in [2]. In Figure 1, we present the original left, right channel images and the disparity estimation and depth map for the Aqua sequence.

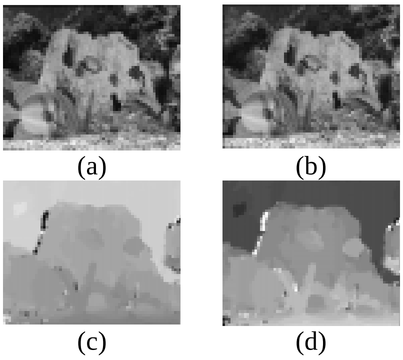

Figure 1. Disparity and depth estimation for the Aqua sequence. (a) Left and (b) right channel image, (c) horizontal disparity field and (d) depth map.

3. Detection and Compensation of Occluded Areas

The above estimation of the disparity assumes that a corresponding point of image I_2 can always be found for all points of image I_1. However, due to interposition of foreground objects there may be areas of I_1 that are occluded in I_2. As a result for every horizontal line segment that is visible by camera 1 (2) and occluded from camera 2 (1), when traversing the segment from left to right, there is a horizontal disparity decrease in I_2 that is equal to the length of the line segment [7].

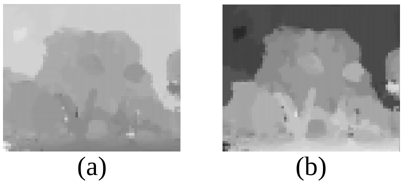

Figure 2. Occlusion detection and compensation for the Aqua sequence. (a) Compensated horizontal disparity field and (b) compensated depth map.

All disparity values obtained through the previous minimization procedure for occluded areas of image I_1 are not reliable and thus may result in incorrect depth segmentation. Therefore, it is clear that (i) these areas should be detected, and (ii) occlusion should be compensated by assigning appropriate disparity values to detected areas. The former task, occlusion detection, is accomplished by locating regions of I_1 where the horizontal disparity decreases continuously with respect to the horizontal coordinate x_1 with a slope approximately equal to –1. The latter task, occlusion compensation, is tackled by keeping disparity constant in each occluded area, and equal to the maximum disparity value of that area. In Figure 2 the occlusion detection and compensation for the Aqua sequence are presented.

4. Video Object Segmentation

The next step of the algorithm is to segment stereo image sequences into semantically meaningful objects. However, semantic video object segmentation is a difficult task with the exception of some specific applications [8]. As we are interested in a fully automatic segmentation algorithm, which is not restricted to specific applications, features computed from the previous analysis, including color and depth information, are used to describe the stereo visual content. In particular, we use the Recursive Shortest Spanning Tree (RSST) algorithm and a segmentation fusion technique, which is presented in the next sub-section. The RRST algorithm was selected because it does not impose any external constraint on the image and also permits simple control over the number of segmented regions [9]. However, the bottleneck of the algorithm is its computational complexity. For this reason, a new multiresolution implementation of the RSST, called M-RSST, is used in this paper, which recursively applies the RSST to images of increasing resolution. This approach, apart from accelerating the segmentation procedure, also reduces the number of small objects, which is a useful property in the context of the proposed video summarization scheme. In Figure 3(a) color segmentation results of the Aqua sequence are presented while in Figure 3(b) we can see depth segmentation results for the same sequence.

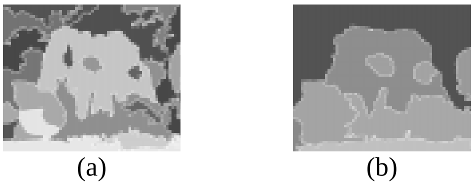

Figure 3. (a) Color segmentation using the M-RSST for the aqua sequence. (b) The respective depth segmentation.

Object boundaries (contours) cannot be identified with high accuracy by a depth segmentation algorithm, due to erroneous estimation of disparity field and occlusion issues. On the contrary, segmentation based on color homogeneity criteria, contains the most reliable object boundaries. Therefore, color and depth segments are appropriately fused together so that a more precise content-based segmentation is accomplished.

Let us assume that K^c color segments and K^d depth segments have been extracted by the aforementioned M-RSST algorithm, denoted as S_i^c, $i=1,2,\ldots,K^c$ and S_i^d , $i=1,2,\ldots,K^d$ respectively. Let us also denote by G^c and G^d the output masks of color and depth segmentation, which are defined as the sets of all color and depth segments respectively:

$$G^c=\{S_i^c, i=1,2,\ldots,K^c\},\quad G^d=\{S_i^d, i=1,2,\ldots,K^d\} \quad (2)$$

Color segments are projected onto depth segments so that video objects provided by depth segmentation are retained and, at the same time, object boundaries given by color segmentation are accurately extracted. For this reason, each color segment S_i^c is associated with a depth segment, so that the area of intersection between the two segments is maximized. This is accomplished by means of a projection function:

$$p(S_i^c, G^d)=\arg\max_{g\in G^d}\{a(g\cap S_i^c)\},\quad i=1,2,\ldots,K^c \qquad (3)$$

where $a(\cdot)$ is the area, i.e., the number of pixels, of a segment. Based on the previous equation, K^d sets of color segments, say C_i, $i=1,2,\ldots,K^d$, are defined, each of which contains all color segments that are projected onto the same depth segment S_i^d:

$$C_i=\{g\in G^c : p(g,G^d)=S_i^d\},\quad i=1,2,\ldots,K^d \qquad (4)$$

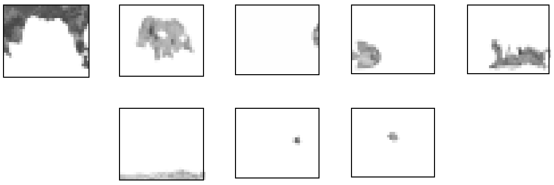

Figure 4. Object extraction after the segmentation fusion for the Aqua sequence.

Then, the final segmentation mask, G, consists of $K=K^d$ segments S_i, $i=1,2,\ldots,K^d$, each of which is generated as the union of all elements of the corresponding set C_i. In Figure 4 the object extraction after the segmentation fusion for the Aqua sequence is shown.

5. Stereo Video Summarization

All features extracted by the stereoscopic video sequence analysis are used to describe the visual content of each video frame. However, they are not directly included in a vector to be used for this purpose, since their size differs between frames. To overcome this problem, we classify color as well as depth segments into pre-determined classes, forming a multidimensional histogram. In this framework, each feature vector element corresponds to a specific feature class (or a bin) and contains the number of segments that belong to this class. In order to reduce the possibility of classifying two similar segments to different classes, causing erroneous comparisons, a degree of membership is allocated to each class, resulting in a *fuzzy classification* formulation [9].

Then, in order to analyze an entire stereoscopic video sequence and extract summarization of its visual content, a shot cut detection algorithm is applied in the beginning. In our approach the algorithm proposed in [10] has been adopted for shot detection due to its efficiency and low computational complexity. Then for every shot we perform key frame extraction. The most characteristic frames for each shot are extracted the ones with the minimum correlation among all the frames of the given shot. For this reason, we define a correlation measure $R_F(\mathbf{a})$ of the frame feature vectors in a shot as

$$R_F(\mathbf{a})=R_F(a_1,\ldots,a_{K_F})=\frac{2}{K_F(K_F-1)}\sum_{i=1}^{K_F-1}\sum_{j=i+1}^{K_F}(\rho_{a_i,a_j})^2 \qquad (5)$$

The vector $\mathbf{a}$ contains the indices of the frames within the examined shot while K_F is the number of selected frames and ρ the respective correlation.

Unfortunately, the complexity of an exhaustive search for the minimum value of $R_F(\mathbf{a})$ is such that a direct implementation would be practically unfeasible. For this reason, a genetic algorithm (GA) [11] approach is adopted. In this approach, possible solutions of the optimization problem, i.e., sets of frames, are represented by chromosomes whose genetic material consists of frame numbers (indices). Chromosomes are thus represented by index vectors following an integer number encoding scheme. An initial population of P chromosomes, $\mathbf{A}(0)=(\mathbf{a}_1,\ldots,\mathbf{a}_P)$ is generated and used for the creation of new generation populations A(n), n>0. The correlation measure $R_F(\mathbf{a})$ is used as an objective function to estimate the performance of all chromosomes $\mathbf{a}_i, i=1,\ldots,P$ in a given population. Then, a *proportionate scheme* is used for parent selection [11], and a set of new chromosomes (offspring) is produced by mating the selected parent chromosomes and applying a *crossover operator*. Finally, *mutation* is applied to the newly created chromosomes, introducing random gene variations that are useful for restoring lost genetic material, or for producing new material that corresponds to new search areas.

Once new chromosomes have been generated for a given population $\mathbf{A}(n),\ \ n \geq 0$, the next generation population, $\mathbf{A}(n+1)$, is formed by inserting those new chromosomes into $\mathbf{A}(n)$ and deleting an appropriate number of older chromosomes, so that each population consists of P members. Several cycles need to take place, that is, several generations $\mathbf{A}(n),\ \ n \geq 0$ need to be produced until the population converges to an optimal solution. Usually the GA terminates when the best chromosome fitness remains constant for a large number of generations, indicating that further optimization is unlikely.

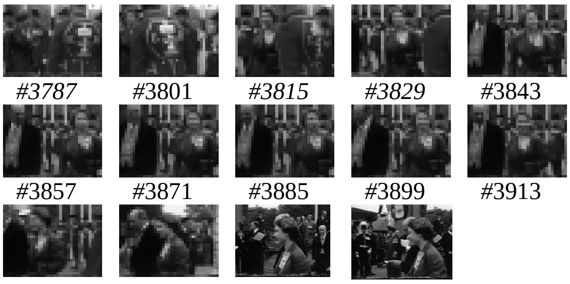

Figure 5. Shot 38 from “Eye to eye” sequence with 88 frames, shown with one frame every 14.

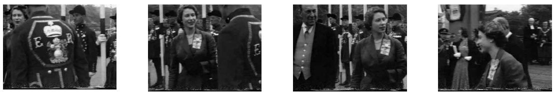

Figure 6. The selected frames of shot 38.

6. Experimental Results

The 3-D stereoscopic television program “Eye to Eye”, of total duration 25 minutes (12,739 frames at 10 frames/sec), has been used in our experiments for the evaluation of the proposed summarization scheme. The stereo video sequence is analyzed according to the described procedure and key-frames are extracted using the genetic algorithm. In Figure 5, shot 38 of the sequence is illustrated, where for presentation purposes, one frame every 7 is shown. The shot consists of 188 frames (stereo pairs) and represents an outdoor crowded scene with considerable camera motion. Furthermore, in Figure 6 we can see the K_F=4 key frames which were extracted by the proposed technique for shot 38.

Finally, Figure 7 presents the minimum value of the correlation measure versus the cycle of the genetic algorithm for shot 38. As expected the $R_F(\mathbf{a})$ decreases as the GA cycle increases, until it reaches a minimum.

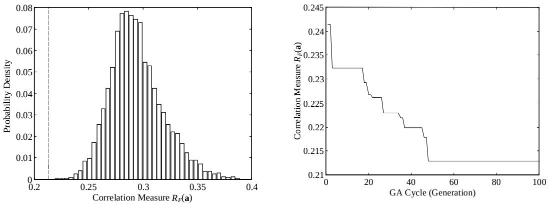

Figure 7. (a) Histogram of correlation measure, together with minimum value obtained from genetic algorithm (vertical dashed line) (b) convergence of genetic algorithm.

7. Acknowledgements

The authors wish to thank very much Professor Chas Girdwood and Professor Siegmund Pastoor for providing the video sequences.

8. References

[1] Special Issue on Content-Based Image Retrieval Systems, *IEEE Computer Magazine*, 28(9), 1995.

[2] D. Tzovaras, N. Grammalidis and M. G. Strintzis, “Disparity Field and Depth Map Coding for Multiview 3D Image Generation,” *Image Communication*, No. 11, pp. 205-230, 1998.

[3] S. W. Smoliar and H. J. Zhang, “Content-Based Video Indexing and Retrieval,” *IEEE Multimedia*, pp.62-72, Summer 1994.

[4] F. Arman, R. Depommier, A. Hsu and M.Y Chiu, “Content-Based Browsing of video Sequences”, *ACM Multim.*, pp. 77-103, Aug. 1994.

[5] N. Vasconcelos and A. Lippman, “A Spatiotemporal Motion Model for Video Summarization,” Proc. of *IEEE CVPR,* pp. 361- 366, Santa Barbara, USA, June 1998.

[6] M. M. Yeung and B.-L. Yeo, “Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content,” *IEEE Trans. Circuits and Systems for Video Techn*, Vol. 7, pp. 771- 785, Oct. 1997.

[7] N. Grammalidis and M. G. Strintzis, “Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multiview Image Sequences,” *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 8, pp. 328-344, June 1998.

[8] R. Castagno, T.Ebrahimi and M. Kunt, “Video Segmentation Based on Multiple Features for Interactive Multimedia Applications,” *IEEE Trans. Circuits and Systems for Video Techn.*, Vol. 8, No. 5, pp. 562-571, 1998.

[9] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, “A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases,” *Computer Vision and Image Understanding*, May 1999 (to appear).

[10] B. L. Yeo and B. Liu, “Rapid Scene Analysis on Compressed Videos,” *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 5, pp. 533- 544, Dec. 1995.

[11] D.E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning, Addison Wesley, 1989.