TEMPORAL SYNCHRONIZATION IN STEREOSCOPIC VIDEO: INFLUENCE ON QUALITY OF EXPERIENCE AND AUTOMATIC ASYNCHRONY DETECTION

Lutz Goldmann, Jong-Seok Lee, Touradj Ebrahimi

Ecole Fédérale Polytechnique de Lausanne (EPFL)
Multimedia Signal Processing Group (MMSPG)
EPFL/STI/IEL/GR-EB, Station 11, 1015 Lausanne, Switzerland

ABSTRACT

In this paper, we analyze the influence of temporal asynchrony on the subjective quality of stereoscopic video. Based on our recently created 3D video database, different levels of asynchrony were simulated and a comprehensive subjective test was conducted to determine the associated degradations in quality of experience. Furthermore, we developed a method to detect asynchrony between left and right video streams based on canonical correlation analysis. Experiments demonstrate the robustness of this method with respect to different amounts of asynchrony and scene depth, which makes it suitable to predict quality of experience or automatic resynchronization.

Index Terms— 3DTV, temporal synchronization, quality of experience

1. INTRODUCTION

The introduction of three dimensional television (3DTV) in the consumer market is believed to be just a matter of time and has been compared to the transition from black-and-white to color TV. To be a success, both visual quality and comfort must be comparable at least to conventional standards to guarantee a strain-free viewing experience. Since 3DTV involves both 2D and 3D visual perception, new distortions have to be considered beside the classical ones.

An important requirement of stereoscopic video is the accurate temporal synchronization between the left and the right camera views. For moving cameras or objects, temporal asynchrony may lead to spatially displaced views, false parallax and ghosting effects that can have a significant impact on the perceived 3D quality. Temporal asynchrony between the two stereo views may be caused by all stages within the overall 3D processing chain. During the acquisition, temporal alignment may be achieved using dedicated synchronization hardware. However, this is only possible for professional applications where expensive equipments are used. Temporal asynchrony may be caused in the video editing stage due to the lack of 3D video editing software which allows the simultaneous editing of both video streams. While 2D video editing software can be used to edit both video streams sequentially, one has to ensure that both streams are manipulated in the same way. For certain applications (e.g. aerial mapping and space telescopy), the two video streams may originate from different hosts and be transmitted independently through wireless networks. Temporal asynchrony may be introduced due to different transmission delays. Finally, temporal asynchrony may be introduced during the restitution due to delays in one of the two decoding processes if the stereoscopic video streams are stored as individual files.

The goal of temporal alignment is to detect and to measure the temporal asynchrony and recover synchronization of the two video streams. Depending on the application and reason for the temporal asynchrony, the analysis may be based on context information (e.g. synchronization markers or presentation timestamps) or the content itself. Content-based resynchronization can be applied for any kind of temporal asynchrony and does not require any additional information apart from the two video streams. Existing methods can be grouped into sequence- and trajectory-based approaches depending on the type of information which is used for the analysis [1]. Sequence-based approaches analyze all the pixels in the video frames. They have the advantage that the spatial transformation between the video streams can be determined more accurately and that they do not require any explicit feature detection or tracking. Trajectory-based techniques rely on feature point tracking and analyze the resulting trajectories. Since trajectory-based approaches contain explicit geometric information, they can cope with large spatio-temporal misalignments, can align video streams acquired with different cameras, and are less affected by background changes.

This paper studies the influence of temporal asynchrony on the perceived 3D quality to derive a range of tolerance (section 2). Furthermore, we propose a robust sequence-based method for temporal asynchrony detection within stereoscopic video streams and evaluate its performance (section 3).

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia).

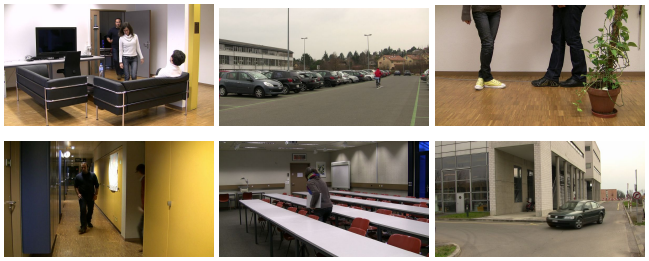

Fig. 1. Visual samples of the scenes used in the experiments. From top left to bottom right: sofa, bike, feet, corridor, notebook, and car.

2. INFLUENCE ON THE SUBJECTIVE QUALITY

2.1. Methodology

2.1.1. Laboratory environment

The subjective test was conducted at the Multimedia Signal Processing Group (MMSPG) quality test laboratory at EPFL, which is compliant with the recommendations for subjective evaluation of visual data issued by ITU-R [2]. A 46" polarized stereoscopic display (Hyundai S465D) with a native resolution of 1920×1080 pixels was used to display the test stimuli. A subject was seated in line with the center of the monitor, at a distance of approximatively 2 m which is equal to 3 times the height of the screen as suggested in the ITU-R BT.500 [2] for HDTV.

2.1.2. Dataset

The experiments are based on a recently created 3D Video Database[1] which contains stereoscopic video sequences of various indoor and outdoor scenes captured with a resolution of 1920×1080 pixels and a frame rate of 25 fps (figure 1).

Only the video sequences with the smallest camera distance (10 cm) from the database were used. The temporal asynchrony was simulated by shifting the stereoscopic video streams frame-wise against each other. Eleven different temporal offsets $\tau = \{$-10, -5, -3, -2, -1, 0, 1, 2, 3, 5, 10$\}$ have been considered, where a negative shift corresponds to a delay of the left video sequence and a positive shift to a delay of the right video sequence.

2.1.3. Procedure

A single stimulus method was adopted for the subjective quality evaluation. In order to determine the influence of the camera distance on the 3D quality, a continuous quality scale with 5 levels (excellent, good, fair, poor, and bad), as described in ITU-R BT.500 [2], was used.

During the training sessions, the subjective test methodology was described to the subjects and the range of quality levels was shown through a set of training stimuli. The stimuli were selected by an expert viewer in such a way that each quality level was represented by an example and that the full range of quality levels within the set of test stimuli was covered.

During the test sessions, subjects evaluated the quality of each of the 66 test stimuli, which were displayed in a random order. Each stimulus was shown once with a duration of 10 s and a 5 s break between the stimuli, during which the subjects provided their scores.

2.1.4. Observers

Twenty subjects (7 female and 13 male) participated in tests. All of them were non-expert viewers with a marginal experience of 3D image and video viewing. The age distribution ranged from 24 to 33 with an average of 27.

2.2. Analysis

2.2.1. Outlier detection

The screening of the subjects was performed according to the guidelines described in ITU-R recommendation [3]. From the 20 subjects 2 were discarded as outliers.

2.2.2. Score computation

After the outlier removal, the mean opinion scores were computed for each stimulus j as

$$MOS_j = \frac{\sum_{i=1}^{N} s_{ij}}{N} \quad (1)$$

where N is the number of considered subjects and s_{ij} is the score from subject i for the test condition j.

Due to the small number of subjects, the confidence intervals (CI) for mean opinion scores were computed as follows:

$$CI_j = t(1-\alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \quad (2)$$

where $t(1-\alpha/2, N)$ is the t-value corresponding to a two-tailed Student's t-distribution with $N-1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). σ_j is the standard deviation of a single test condition across the subjects.

2.3. Results

Figure 2 summarizes the subjective test results by showing the subjective quality scores versus the temporal shift for 4 of the 6 scenes. Based on the observation in [4] that non-expert viewers can only distinguish a few quality levels, the range of quality scores is split into 3 parts (good, fair, and bad) by the horizontal lines in the graphs.

[1] http://mmspg.epfl.ch/3dvqa

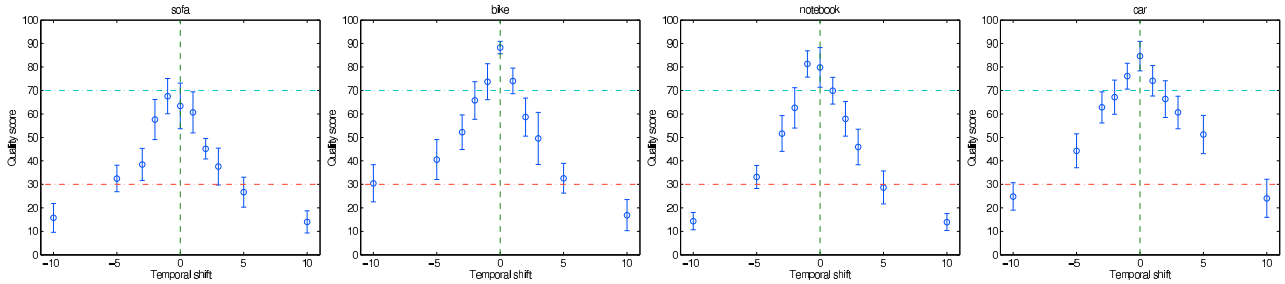

Fig. 2. Mean opinion scores and confidence intervals vs. temporal asynchrony for different scenes (sofa, bike, notebook, car).

In general, the small confidence intervals show that the complexity of the subjective evaluation tasks was appropriate and that the ratings are quite consistent across subjects. Apart from "sofa, all the curves show a similar trend. "Excellent" quality is achieved only for a temporal shift $\tau = 0$. Depending on the scene, the quality is "good" for $\tau \leq 2$. For $\tau > 2$, the quality degrades rapidly and is considered "bad" for $\tau > 5$. For "sofa", the quality for $\tau = 0$ is even below the "good" range. Another interesting observation is the asymmetry of the subjective quality with respect to the direction of the temporal shift. For all scenes except for "car", a negative temporal shift has a smaller influence on the subjective quality when compared to a positive temporal shift of the same amount. This effect might be caused by the direction of the dominant object motion and the decreasing or increasing disparity depending on the direction of the temporal shift. While the horizontal motion in the "car" scene is mainly from right to left, all the other scenes contain left-to-right motions.

3. AUTOMATIC ASYNCHRONY ESTIMATION

3.1. Algorithm

In this section, we present our proposed algorithm for automatic asynchrony detection of stereoscopic video sequences. Basically, the algorithm analyzes the correlation between the visual motion information in the right and the left video sequences in order to estimate the temporal offset τ between them. In other words, by shifting the right video sequence by t frames relative to the left, it searches for the shift value $\hat{\tau}$ producing the maximal correlation. Especially, in order to consider the spatial disparity in the left and right images, the method includes the procedure of spatial matching between different regions of the two images.

In our method, the temporal difference of the luminance component is used in order to obtain the motion information in the scene. For a fixed shift value t, the whole sequence is divided into temporal blocks which are the basic units of the correlation analysis. For each block, the correlation analysis is performed as follows: First, each differential image of each view is divided into small spatio-temporal tiles. Then, the correlations between the tiles in the two views are measured by using the canonical correlation analysis (CCA). Using CCA is motivated by the fact that the disparities between the two views may vary within a tile depending on the horizontal pixel location. For each tile in a left image, a few tiles in the corresponding right image are selected as candidates of matching tiles. Assuming that the two views are vertically aligned, for the tile in the i-th column in the left image we choose the tiles of the column indices between i and $i + m - 1$ in the same row in the right image, where $m > 0$ is the maximum number of candidate tiles to be considered for matching. The correlation between the current tile of the left image and each candidate of the right image is calculated and the maximum value is chosen.

The CCA aims at finding the projection vectors by which the correlation of the projected data becomes maximal. If we let $\mathbf{x}$ and $\mathbf{y}$ be the two vectors containing the mean-normalized differential pixel values in the tile of the left and the right views, respectively, the canonical correlation between them is given by

$$C(\mathbf{x}, \mathbf{y}) = \arg \max_{\mathbf{w_x}, \mathbf{w_y}} \frac{E\{(\mathbf{w_x}^T \mathbf{x})(\mathbf{w_y}^T \mathbf{y})\}}{\sqrt{E\{\mathbf{w_x}^T \mathbf{x}\}^2 E\{\mathbf{w_y}^T \mathbf{y}\}^2}} \quad (3)$$

where $\mathbf{w_x}$ and $\mathbf{w_y}$ are the projection vectors. This maximization problem can be resolved by solving an eigenvalue problem [5]. In order to obtain a unique solution, it is necessary to choose the length of the temporal blocks to be larger than the length of $\mathbf{x}$ or $\mathbf{y}$.

The obtained canonical correlation values for all tiles in the left image are averaged. The resulting correlation values are averaged again over the temporal blocks in order to obtain the overall correlation measure for the current temporal shift t. Finally, after we calculate the correlation measures for all temporal shift values, the estimated offset $\hat{\tau}$ is obtained by finding the maximum correlation measure.

3.2. Experiments

We use the same stereoscopic sequences as in the subjective tests described in Section 2.1.2. The image frames were downscaled to 1/64 of their original size in order to reduce computational complexity. Each frame was divided into 4×4 tiles for analysis. We used temporal blocks with length of 50 frames and an overlap of 50% with the precedent and the sub-

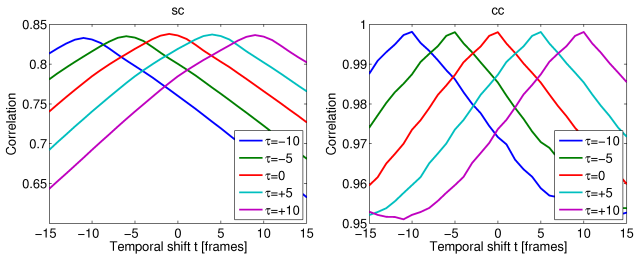

Fig. 3. Measured correlation values by the simple correlation (left) and the canonical correlation methods (right) for "feet".

Method	sofa	bike	feet	corr.	noteb.	car	All
SC	0.73	0.00	1.00	0.00	0.00	1.00	0.45
CC	0.55	0.00	0.00	0.00	0.00	0.55	0.18

Table 1. Error rates of the simple correlation (SC) and canonical correlation (CC) methods for asynchrony detection.

sequent ones. The search range of the temporal offset was set to $[-15, -14, \cdots, 15]$.

In order to evaluate the performance of the proposed method, we also ran a simple method based on the frame-by-frame correlation measure as a baseline. In other words, by temporally shifting the right image within the search range, the correlation coefficient is calculated for the pixel values of the image pairs of the two views at each time frame, which is averaged over time. This method does not consider the spatial matching of the two views having disparity and can be considered as the simplest way of estimating asynchrony based on correlation analysis.

Figure 3 illustrates the measured correlation values by the two methods with the shift values in the search range for five different given offsets. The correlation value becomes the maximum when the shift value matches the given offset, and decreases as their difference becomes large. It is observed that, while the simple method produces an error of one frame for all test offsets, the proposed method estimates the shift correctly for all cases, as shown in Table 1. In addition, the correlation values by the simple method are lower than those by the proposed method because CCA maximizes the correlation by projecting the data from the two views.

Table 1 compares the performance of the simple correlation (SC) and canonical correlation (CC) methods for all test sequences. Since the maximum amount of the errors in both methods was one frame, the table shows the frequency of the error among 11 test offsets. Overall, it is observed that the proposed method outperforms the simple correlation method since the latter is quite sensitive to variations in the 3D scene structure and dominant object motion.

The result of the horizontal spatial matching of the two views by the proposed method is illustrated in Fig. 4, which shows the horizontal index of the matching tile of the right view for each tile of the left view for the first temporal block

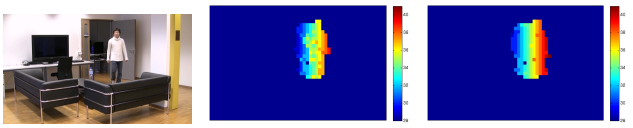

Fig. 4. Original left image (left) and horizontal tile matching results when the two views are not aligned (middle) and are aligned (right) for "sofa". The horizontal index of the matching tile of the right image for each tile in the left image is shown with different colors in the middle and right figures. Only the moving object (the waking person) was considered for matching.

of "sofa". When the two views are temporally aligned (right figure), two kinds of consistency are preserved better than when the views are not aligned (middle figure): (1) the matching tile index increases monotonically along the horizontal direction for almost all tiles; (2) the matching tile index is mostly constant along each column of the image.

4. CONCLUSION

We have studied the influence of temporal asynchrony on the subjective quality. It has been shown that an offset of more than 2 frames leads to a significant quality drop. In order to detect asynchrony of the two video streams, a method for automatic asynchrony estimation has been developed based on a combination of block-wise disparity estimation and CCA. It has been shown to perform successfully for various scenes with considerable improvement in comparison to simple correlation. Overall, the resynchronization by the two methods will achieve "good" subjective quality for various contents and given temporal shifts with a maximum error of one frame.

5. REFERENCES

[1] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood, "View invariant alignment and matching of video sequences," in *Proc. ICCV*, 2003, pp. 939–945.

[2] ITU-R, "Subjective assessment of stereoscopic television pictures," Tech. Rep. BT.1438, 2000.

[3] ITU-R, "Methodology for the subjective assessment of video quality in multimedia applications," Tech. Rep. BT.1788, ITU-R, 2007.

[4] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *EI, 3DIP*, 2010.

[5] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, pp. 2639–2664, 2004.