Rapid 2D-3D Conversion for Low-Cost 3D Television

Tamer Rabie
Associate Professor of Computer Engineering
Department of Electrical & Computer Engineering
University of Sharjah, U.A.E.
trabie@sharjah.ac.ae

ABSTRACT

This work develops a real-time 2D-to-3D converter that exploits motion parallax naturally available in a normal 2D moving image sequence to produce a 3D side-by-side motion picture suitable for viewing on low-cost 3D Television displays at conversion processing rates that can reach high speeds of up to 100 frames per second. The novel paradigm presented in this paper enhances 3D perceptibility by ensuring continuous synchronization between the left and the right image views even if motion in the 2D video abruptly freezes or transitions rapidly, with a minor single frame initial 2D broadcast, equivalent to an initial 2D delay of 1/30 of a second for a 30 frames per second video stream, before full 3D takes effect, and neither depends on computationally expensive depth map extraction nor does it require any special hardware setup such as multicore processors or special purpose graphics processing units.

General Terms:

3D Visualization, 3D Multimedia, Human Stereo Perception.

Keywords:

2D-3D Converter, Motion Parallax, 3DTV, Stereo Vision.

1. INTRODUCTION

The past decade has seen a plethora of affordable high-definition (HD) 3D television (3DTV) displays being made available to a wide spectrum of users. However, the high cost of commercial HD 3D broadcast transmission has restricted the user from making full use of this new technology by limiting its utility to viewing 3D content either offline by playing native 3D BlueRay movies, or through the relatively poor quality 2D-to-3D (2D-3D) converters built into most 3DTV sets for the purpose of converting the 2D broadcast transmission to 3D on-the-fly.

2D-3D conversion adds the binocular disparity depth cue to digital images perceived by the brain, thus greatly improving the immersive effect while viewing stereo video in comparison to single-lense 2D video. However, in order to be successful, the conversion should be done with sufficient accuracy and correctness: the quality of the original 2D images should not deteriorate, and the introduced disparity cue should not contradict to other cues used by the brain for depth perception. If done properly and thoroughly, the conversion produces stereo video of similar quality to ”native” stereo video acquired using a dual-lense stereo camera and accurately adjusted and aligned in post-production [1].

Fig. 1. Left: One frame from a scene of the movie “Back to the future”. Right: Generated depth map of objects in this scene (Courtesy Triaxes http://triaxes.com). Extracting dense depth maps from 2D scenes is extremely computationally intensive which makes rendering 2D-3D movies in real time almost impossible.

Two approaches to stereo conversion can be loosely defined: quality semiautomatic conversion for cinema and high quality 3DTV, and low-quality automatic conversion for low-cost 3DTV, Video on Demand (VoD)[1] and similar applications.

Recent attempts to improve the quality of 2D-3D video conversion has been discussed in the literature [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. A shortcoming that is common among these techniques is their dependability on computationally expensive depth-map algorithms for depth information extraction and depth image-based rendering thus requiring expensive hardware such as multicore processors and special purpose graphics processing units to perform the conversion of the 2D video content to a stereoscopic 3D version in real-time, which directly affects the affordability of the 3DTV display that encorporated this type of hardware/software setup. An example frame from the motion picture “Back to the future” and its corresponding depth map calculated for objects in the scene is shown in figure 1.

In contrast, the focus in this work is on real-time performance for low-cost 3DTV hardware. This places some constrains on processing power thus prompting us to take a simpler

[1] Video on demand (VoD) is an interactive TV technology that allows subscribers to view programming in real time or download programs and view them later.

approach to the problem of 2D-3D conversion while sustaining visually comfortable 3D depth perception throughout the broadcast trasmission.

Most TV content is based on dynamic motion-picture scenes. Rarely would a TV channel broadcast a still picture all the time, and even if it did, converting it to 3D would be of little benefit to the viewer. Thus, the demand on 3D is typically confined to real-time motion picture content. The depth cue that is most readily available in a motion picture is motion parallax.

Motion parallax is a natural phenomena that provides monocular depth cues to the human visual system allowing perception of relative object distances in the environment when an observer closes one eye and looks with the other [15]. When either an observer or the objects in a scene move, the apparent relative motion of objects with respect to the observer give hints about their relative distance. If information about the direction and velocity of movement is known, motion parallax can become a powerful cue to provide absolute depth information [16]. This effect can be seen clearly when driving in a car. Nearby objects pass quickly, while distant objects appear to move much slower. Some animals that lack binocular vision due to their eyes having little common field-of-view employ motion parallax more explicitly than humans for depth cueing.

Motion parallax has been widely used in the literature to extract depth information from monocular video sequences acquired using single-lense cameras. Algorithms such as structure-from-stereo [17, 18, 19] and structure-from-motion [20, 21, 22] can be applied for this purpose. Those algorithms consist of two parts: determination of motion parallax from the sequence, and the mapping of motion parallax into depth information. Although these algorithms provide effective depth information to generate depth maps for the objects in the scene, they are typically computationally expensive to implement and deploy in applications related to generation of 3D from 2D motion pictures.

The work described in this paper exploits the natural phenomenon of motion parallax in monocular single-lense video and develops a relatively simple paradigm to convert monocular 2D video sequences to binocular 3D side-by-side (SBS) video sequences of the same length in real-time making use of the inherent depth perception from our binocular human visual systems to generate visually acceptable 3D information, while handling degenerate cases where other simple 2D-3D converters fail.

The rest of this paper is organized as follows. In section 2, a brief review of how the human stereo visual system develops the sense of depth from binocular vision is presented. Section 3 gives a detailed account of the proposed high-speed 2D-3D conversion method and also presents some results which demonstrate the effectiveness of the proposed algorithm in handling rapid object motions which causes other methods to fail. Finally, concluding remarks appear in section 4.

2. THE HUMAN BINOCULAR VISUAL SYSTEM

To understand how motion parallax can be used effectively to reproduce 3D depth information from 2D motion pictures, one must study how the human stereo visual system develops the sense of depth from a pair of eyes.

Stereo vision refers to the sense of depth that is perceived when a scene is viewed with both eyes by someone with normal binocular vision. Binocular viewing of a scene creates two slightly different image projections of the scene objects in the retina of the two eyes due to the eyes' different positions on the head. These differences, referred to as binocular stereo disparity, provide information that the brain can use to calculate depth in the visual scene. It is clear from figure 2 that disparity varies with the proximity of the object from the eyes, increasing as the object comes closer and vise versa. Once the human eyes are verged on an object, it is straightforward for the human brain to sense the range to the object. This depth information can also be estimated mathematically from the gaze angles. Referring to Fig. 3, the depth to the object is given by [23]:

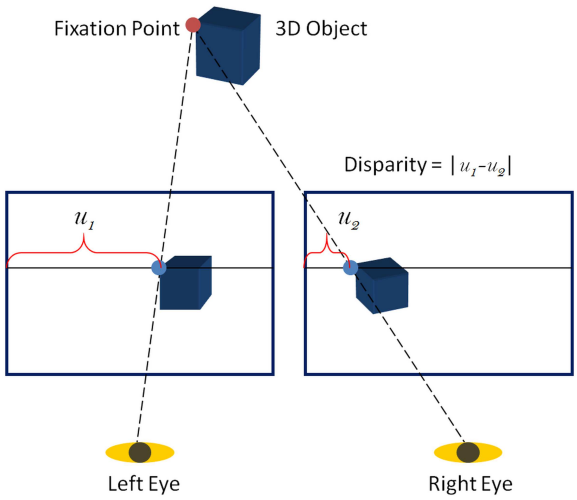

Fig. 2. Binocular viewing of a scene creates two slightly different images of the scene in the two eyes due to the eyes' different positions on the head. These differences, referred to as disparity, provide information that the brain can use to calculate object depth in the visual scene.

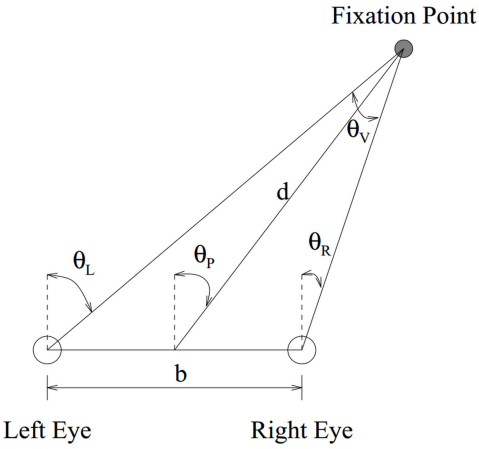

Fig. 3. Gaze angles and range to object geometry for a typical Human Stereo Visual System.

$$d = b\frac{\cos(\theta_R)\cos(\theta_L)}{\sin(\theta_L - \theta_R)\cos(\theta_P)}, \quad (1)$$

where b is the baseline between the two eyes, and $\theta_P = \frac{1}{2}(\theta_R + \theta_L)$. When the eyes are verged on an object the vergence angle is $\theta_V = (\theta_R - \theta_L)$ and its magnitude increases as the object comes closer to the eyes [24].

3. METHODS AND RESULTS

The method developed in this work implements an incremental low-computational-cost 2D-3D video conversion paradigm for real-time 2D TV broadcasts viewed on 3DTV displays. The method adopts a coarse-to-fine, level-of-detail (LOD) approach that varies the strength of perceived motion parallax depending on the moving object's detail level, which allows for a more immersive experience. High-detailed objects are closer to the camera and thus have large motions and stronger motion parallax effect, while low-detailed objects are farther away from the camera and have smaller motions and consequently weaker motion parallax.

The idea is to adaptively vary a motion threshold depending on the LOD of objects in the scene between consequtive frames. The larger the motions the higher the motion parallax and in this case only adjacent frames are used to render the SBS 3D frame. If the closer objects disappear from the scene and only slow moving objects with less details are present (usually in the background), then the second to previous or third to previous frame is used with the current frame to render the SBS 3D frame.

This technique, nevertheless, only requires continuous storage of the previous three frames of the running motion picture at any given time, thus memory overhead requirements are minimal. The method also eschews computationally expensive optical-flow-based depth maps and depth image-based rendering techniques, common among most 2D-3D conversion techniques, in favour of simpler methods for estimating a measure of pixel displacements between consecutive frames.

Figure 4 shows a block diagram of the 2D-3D continuous conversion paradigm. The merits of this method is that it is fully automated (i.e. does not require any intervention from the user) requiring a single fixed global threshold value ($T_2 = 0.001$) which is built-in the algorithm and adaptively altered depending on the type of motion sequences being processed. This is unlike other semi-automatic techniques which require users to supply input parameters and to manually alter them.

Referring to figure 4, when the 2D-3D algorithm is started, with the first frame given by $f(i-3)$, it will initially generate a left/right (SBS) stereo frame depicted by $f_{sbs}(i-3)$ in the block diagram. This is only a 2D frame due to lack of any motion parallax to allow a 3D depth perception. Next, the second incoming frame given by $f(i-2)$ is combined with $f(i-3)$ to form the next SBS 3D frame $f_{sbs}(i-2)$. The next frame $f(i-1)$ is then compared to its previous frame $f(i-2)$ using the well known mean-square-error (MSE) metric given by:

$$MSE_L = \frac{1}{M \times N} \sum_{s=0}^{(M \times N)-1} |f_i(s) - f_{i-L}(s)|^2, \qquad (2)$$

for image sizes of $M \times N$, where $L = [1, 2, 3]$ is the current level of detail being checked.

MSE calculates the global error variance (power in the difference image) between two images $|f_i - f_{i-L}|$ and has been widely used for measuring the performance of various image processing filters [25]. The merit in using the MSE_L metric is that it is sensitive to minor pixel variations between the two images which is important for slow moving low-detail objects.

In the degenerate case where motion of objects rapidly jump between consecutive frames, surpassing the natural spatial disparity values for normal human stereo vision perception, both the left and right images are set to the right frame thus reducing the 3D effect to a normal 2D picture. As mentioned, this is a degenrate case and does not occur too often in real-world motion pictures. One example is shown in figure 5 where it was not possible to use two consecutive frames to create the 3D picture due to the huge jump in the motion of the baseball pitcher between these two frames. This technique is also applied to the frequently occuring degenerate case of a transition between scenes in a motion picture where both left and right images are set to the right image, thus momentarily reducing the 3D effect to a 2D picture of the first frame in the new scene.

For both degenerate cases if the MSE of pixel displacement between the newly generated left and right image frame is found to be larger than a threshold $T_1 = 10 \times T_2$, then both left and right images are set to the right image, thus momentarily reducing the 3D effect to a 2D picture. The 2D-3D conversion then restarts as if applying the algorithm to a broadcast transmission for the first time. This technique prevents eye strain from an incorrect transition of scenes where the right image is the new scene image and the left image is still an old scene image from a couple of frames back, and also prevents an unsynchronized left/right image pair due to rapidly jumping object motion between frames, as is clear in figure 5.

The value of T_1 is adaptively chosen by the algorithm. Typically MSE values for large scene motions was found to be within the range $[0.01 \cdots 0.05]$ depending on the value of the largest temporal disparity of moving objects in the scene, while typical scene transition MSE values jump to around $MSE >= 0.15$ which is why the same T_1 threshold can be used to prevent both degenerate cases. For example, scenes with rapidly moving objects require T_1 to be in the upper limits towards the 0.05 range while for scenes with slower moving objects T_1 is towards the lower end values. This way faster moving objects will not be mistaken for a change in scenes between frames and 2D-3D conversion will continue to take effect. Figure 6 clearly shows this variation in the MSE values for both rapidly moving frames (frames 33 to 49), and scene transition frames (between frame 101/102 and between frames 331/332).

The technique used to adaptively choose a proper value for threshold T_1 is to start with the lower end $T_1 = 0.01$ and then calculate the MSE_1 of temporal disparity between current frame and the immediate previous frame. If this value is found to be less than the current T_1 value, indicating slower moving objects in the scene, then the value of T_1 is left unchanged and used for testing the two degenerate cases mentioned above. Otherwise the value of T_1 is multiplied by a constant to anticipate larger motions in the next frame; $T_1 = \tau T_1$, where τ is allowed to cycle through $[1.5 \cdots 1.8]$, after which T_1 is reset to the lower end value (0.01) and τ is reset to 1.5. Once a degenerate case appears, the MSE value will be greater than T_1 and the corrective action explained above is taken, after which T_1 and τ are again reset back to their lower end values respectively.

Figure 7 presents example 3D frames rendered by the proposed 2D-3D method from 2D motion pictures showing high quality conversions. The top right stereo SBS frames include depth of field (DOF) blurring for distant slow moving pixels and no blurring for closer fast moving pixels (the house on the left) which may enhance the 3D immersive perception. DOF was implemented by simply comparing pixel values of the absolute difference images $|f_i - f_{i-L}|$, used to compute the MSE_L given in equation 2, to the adaptive threshold T_1, and blurring those pixels that fall below this threshold (indicating slower moving pixels representing distant pixels). To compare to non-DOF 3D refer to the stereo SBS pair immediately below the top-right. The proposed method is also effective in converting old black-and-white film as is shown by the stereo SBS frame pair on the left. The bottom rows of the figure show corresponding 3D anaglyph versions of the proposed 2D-3D

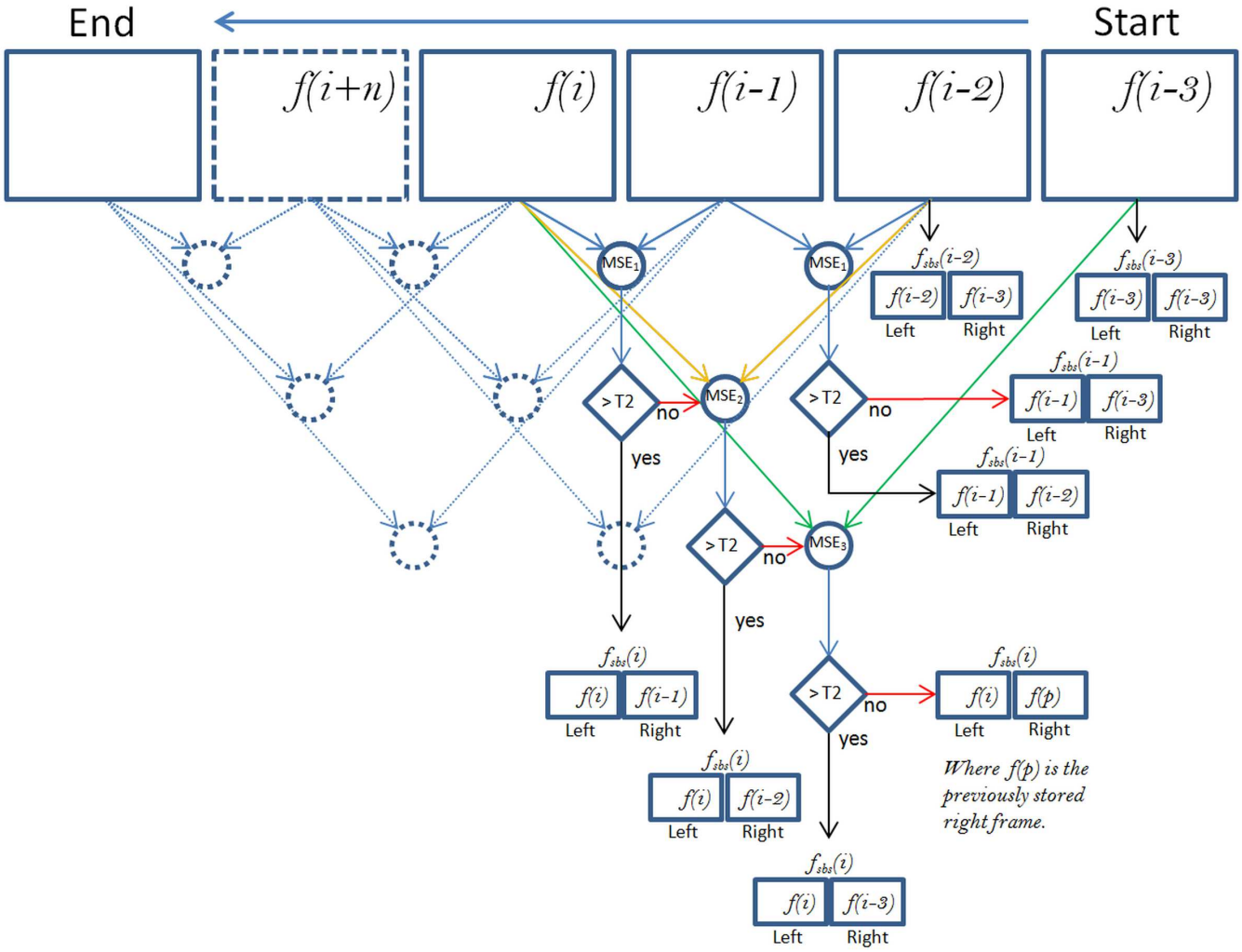

Fig. 4. Block diagram of 2D-3D system developed to generate continous 3D from monocular motion picture transmission.

conversions shown in the top rows, including examples converted from the 2D motion picture "Gravity" and the 2D LG demo movie, to be viewed using red/cyan stereo glasses.

Lastly, the results of a modification in the proposed 2D-3D conversion algorithm to handle the special case of extreme slow motion video is presented. By setting the maximum allowable LOD to be $L = [1 \cdots 7]$ instead of $L = [1, 2, 3]$, the algorithm is allowed to compare the current frame f_i with up to 7 previous frames $f_{i-L}; L = [1 \cdots 7]$ using the MSE_L metric defined in equation 2. The algorithm then continues, as depicted in figure 4, but with 7 levels instead of 3 (calculating $MSE_1, MSE_2, \cdots, MSE_7$ as needed). Decision to increase the number of levels is taken automatically when the algorithm detects extreme slow motion between frames in the current running scene sequence; indicated by an initially small $MSE_1 < \frac{T_2}{10}$ value. Increasing the number of levels increases the memory overhead to 7 previous stored frames, and also slightly increases the processing time which does not affect real-time performance due to simplicity of the calculations, but allows for more motion parallax to be included and consequently an increased immersive 3D perceptual effect.

The top row of figure 8 shows a 3D stereo SBS frame pair converted using the modified 2D-3D method applied to a 2D movie that was extracted from the left frames of an extremely slow motion portion of the LG 3D Demo movie. The LOD was automatically set to the maximum value of 7, due to extreme slow motion detected, which allowed the stereo SBS frames to be rendered 7 frames apart. The top-right image is the corresponding 3D anaglyph version of the SBS stereo pair to be viewed using red/cyan stereo glasses. The bottom stereo SBS pair is the equivalent stereo frame extracted from the native 3D LG Demo movie, and next to it on the bottom-right is the corresponding 3D anaglyph version. It is clear that results rendered using the proposed technique are close to the native 3D LG Demo frame acquired using a native 3D camera.

4. CONCLUSIONS

The author has presented a framework for high-quality 2D-3D conversion from monocular motion pictures of variable fidelity. The idea is to utilize the natural phenomenon of motion parallax in monocular single-lense video and develop a simple paradigm that can be deployed in low-cost 3DTVs to convert monocular

Fig. 5. Degenerate case examples for LG 3D Demo video: (Top Row) Frames 44 to 49 have objects that are rapidly jumping between frames thus preventing proper 3D generation. (Second Row) The proposed 2D-3D method has correctly synchronized the SBS 3D frames rendered from 2D top row frames showing lack of 3D in frames 45 to 48 due to rapidly jumping motions in foreground object (the baseball pitcher) with 3D starting to become available from frame 49 due to slowing down of object motion to normal speed. (Third Row) The original native 3D frames extracted from LG stereo 3D demo movie shown for comparison (3D SBS demo video is publicly made available on the web courtesy of LG). (Lower Row) Frames 46 to 49 converted to 3D by the iQmango commercial 3D player showing erroneous 3D conversion for the rapidly moving pitcher frames.

2D broadcast transmission to binocular 3D SBS video sequences of the same length in real-time making use of the inherent depth perception from the binocular human visual systems to generate visually acceptable 3D information. The paradigm is also able to handle degenerate cases where other simple 2D-3D conversion algorithms fail. Comparison results have shown that the proposed methodology is overall as good as native 3D stereo video with the exception of degenerate cases which are handled properly to minimize the unpleasant effects of these otherwise unsynchronized stereo frames.

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers for their useful comments which helped improve the presentation of this paper. This work was funded by the College of Graduate Studies and Research at the University of Sharjah under project number 140440 for 2014.

5. REFERENCES

[1] B. Sandrew, “2D-3D conversion can be better than native 3D,” http://www.3dfocus.co.uk/3d-features/2d-3d-conversion-interview-legend-3d-barry-sandrew/1394, January 2011.

[2] H. Murata, Y. Mori, S. Yamashita, A. Maenaka, S. Okada, K. Oyamada, and S. Kishimoto, “A real-time 2-D to 3-D image conversion technique using computed image depth,” in *SID Symposium Digest of Technical Papers*, vol. 29. Wiley Online Library, 1998, pp. 919–923.

[3] C. Fehn, P. Kauff, M. O. De Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, “An evolutionary and optimised approach on 3D-TV,” in *Proc. of IBC*, vol. 2, 2002, pp. 357–365.

[4] P. V. Harman, J. Flack, S. Fox, and M. Dowley, “Rapid 2D-to-3D conversion,” in *Electronic Imaging 2002*. International Society for Optics and Photonics, 2002, pp. 78–86.

[5] C. Fehn, “Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3D-TV,” in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 93–104.

[6] I. Ideses, L. P. Yaroslavsky, and B. Fishbain, “Real-time 2D to 3D video conversion,” *Journal of Real-Time Image Processing*, vol. 2, no. 1, pp. 3–9, 2007.

[7] Y.-L. Chang, C.-Y. Fang, L.-F. Ding, S.-Y. Chen, and L.-G. Chen, “Depth map generation for 2D-to-3D conversion by short-term motion assisted color segmentation,” in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1958–1961.

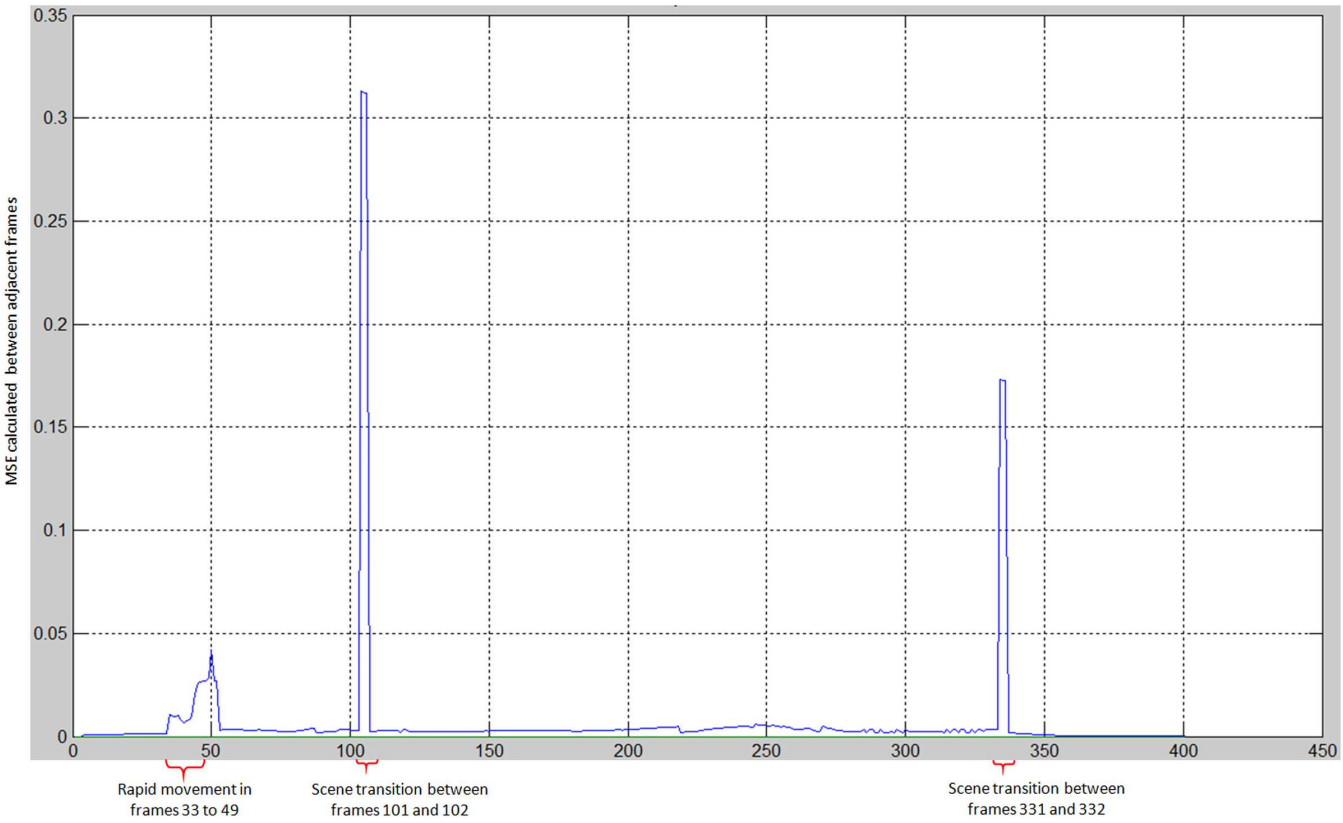

Fig. 6. Plot of frame numbers versus MSE values calculated between consecutive frames of a segment of the 2D LG demo movie (inclusive of frames 44 to 50 shown in figure 5). It is clear that during rapidly moving objects (frames 33 to 49) and for scene transition frames (between frame 101/102 and between frames 331/332) the MSE values jump to large values above the value of T_1 threshold thus prompting the proposed algorithm to deal with these degenerate cases as described in the text.

[8] M. T. Pourazad, P. Nasiopoulos, and R. K. Ward, "An h. 264-based scheme for 2D to 3D video conversion," *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 2, pp. 742–748, 2009.

[9] K. Yamada and Y. Suzuki, "Real-time 2D-to-3D conversion at full hd 1080p resolution," in *Consumer Electronics, 2009. ISCE'09. IEEE 13th International Symposium on*. IEEE, 2009, pp. 103–106.

[10] K. Fliegel, "Advances in 3D imaging systems: Are you ready to buy a new 3D TV set?" in *Radioelektronika (RADIOELEKTRONIKA), 2010 20th International Conference*. IEEE, 2010, pp. 1–6.

[11] B. Coll, F. Ishtiaq, and K. OConnell, "3D TV at home: Status, challenges and solutions for delivering a high quality experience," in *Int. Workshop Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[12] S.-F. Tsai, C.-C. Cheng, C.-T. Li, and L.-G. Chen, "A real-time 1080p 2D-to-3D video conversion system," *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 2, pp. 915–922, 2011.

[13] G. Bravo, L. Do, S. Zinger, and P. de With, "Real-time free-viewpoint DIBR on GPUs for 3DTV systems," in *Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–4.

[14] Y.-K. Lai, Y.-F. Lai, and Y.-C. Chen, "An effective hybrid depth-generation algorithm for 2D-to-3D conversion in 3D displays," *Display Technology, Journal of*, vol. 9, no. 3, pp. 154–161, 2013.

[15] H. E. Burton, "The optics of euclid," *JOSA*, vol. 35, no. 5, pp. 357–357, 1945.

[16] S. H. Ferris, "Motion parallax and absolute distance." *Journal of experimental psychology*, vol. 95, no. 2, p. 258, 1972.

[17] E. Arce and J. L. Marroquin, "High-precision stereo disparity estimation using hmmf models," *Image and Vision Computing*, vol. 25, no. 5, pp. 623–636, 2007.

[18] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 3, pp. 321–334, 2004.

[19] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[20] T. Jebara, A. Azarbayejani, and A. Pentland, "3D structure from 2D motion," *Signal Processing Magazine, IEEE*, vol. 16, no. 3, pp. 66–84, 1999.

[21] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

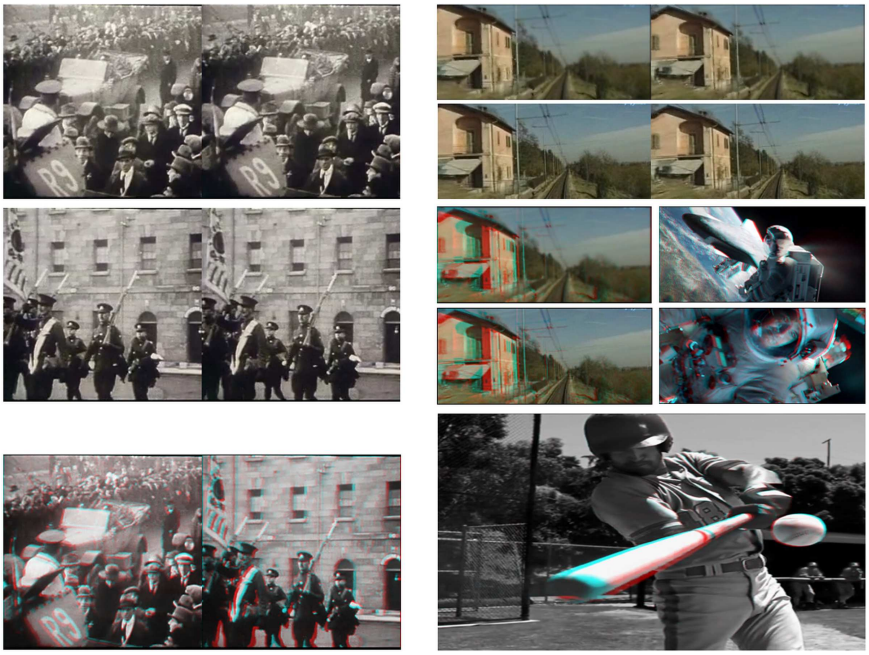

Fig. 7. Example 3D frames rendered by the proposed 2D-3D method from 2D motion pictures showing high quality conversions. Top rows: SBS 3D stereo pairs, Bottom row: corresponding 3D anaglyph version of the SBS stereo pairs to be viewed using red/cyan stereo glasses.

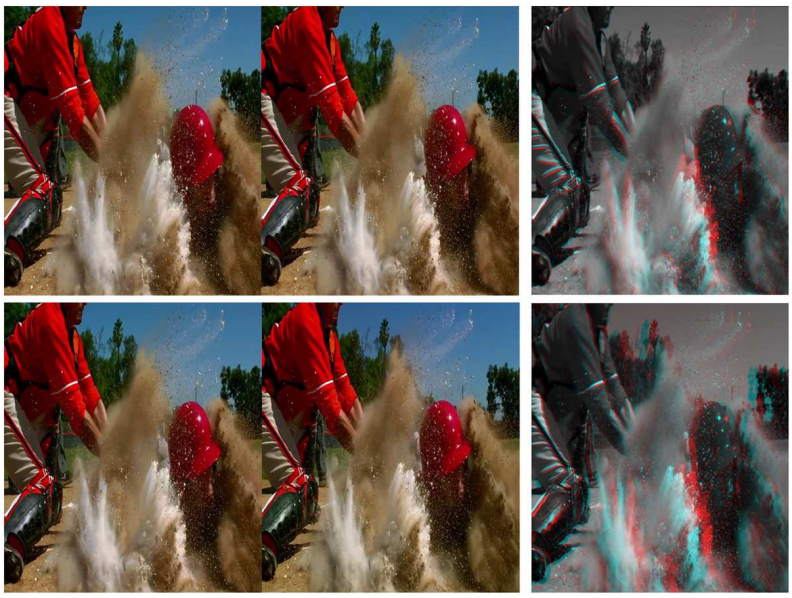

Fig. 8. A comparison between the proposed 2D-3D conversion method modified for detecting extreme slow motion, and a native 3D stereo frame from the LG 3D Demo movie. The top stereo frame pair show 3D results rendered using the proposed method converted from an extreme slow motion scene in the LG 2D movie (extracted from the left frames of the LG 3D Demo movie). The bottom stereo pair is the equivalent stereo 3D frame from the native 3D LG Demo movie. It is clear that results rendered using the proposed technique are close to the native 3D LG Demo frame acquired using a native 3D camera.

[22] E. Imre, A. Alatan, S. Knorr, and T. Sikora, "Prioritized sequential 3D reconstruction in video sequences of dynamic scenes," in *Signal Processing and Communications Applications, 2006 IEEE 14th*. IEEE, 2006, pp. 1–4.

[23] B. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1986.

[24] D. Coombs and C. Brown, "Real-time binocular smooth pursuit," *Inter. J. Computer Vision*, vol. 11, no. 2, pp. 147–164, 1993.

[25] K. Castleman, *Digital Image Processing*. Upper Saddle, NJ: Prentice Hall, 1996.