A TECHNIQUE FOR FAKE 3D (2D-TO-3D CONVERTED) VIDEO RECOGNITION

Efstratios Kakaletsis, Nikos Nikolaidis *

Artificial Intelligence & Information Analysis Lab
Department of Informatics
Aristotle University of Thessaloniki
Box 451, GR-54124 Thessaloniki, GREECE
Emails: *ekakalet@aiia.csd.auth.gr, nikolaid@aiia.csd.auth.gr*

ABSTRACT

In this paper, we propose a technique for the automatic recognition of "fake" stereoscopic videos/movies i.e., videos which result from classic 2D videos through a 2D to 3D conversion process. Essentially, the proposed technique distinguishes between 2D movies converted to 3D and real stereoscopic ones. It is based on the difference in sharpness around foreground objects in a converted stereo frame pair caused from the inpainting step that takes place after the generation of the right frame (rendered view) from the left frame (source view). The two variants of the algorithm, one utilizing a two-class Support Vector Machine and another one that follows a threshold based classification approach, use a sharpness metric evaluated on a stripe created around foreground objects such as human figures. Experimental evaluation of the proposed algorithm, which can serve as 3D quality characterization tool, is conducted on several stereoscopic movies with very promising results.

Index Terms— 3DTV, 3D cinema, stereoscopic video, Real 3D video, Fake 3D video, 2D to 3D Video Conversion, quality assessment

1. INTRODUCTION

The introduction of 3D cinema and television (3DTV) in recent years led to a substantial demand for stereoscopic material such as 3D movies and 3D TV programmes. Due to the high cost of the creation of "real" 3D stereoscopic material i.e., material created with the use of stereoscopic video cameras and 3D rigs, the production of 3D video with 2D-to-3D conversion methods, namely through the post processing of 2D material, has received considerable attention. Both automatic and manual or semi-automatic conversion methods exist. The main steps of automatic 2D-to-3D video conversion are the extraction of the depth information of the scene and the creation of the stereoscopic 3D video with Image 3D Warping or DIBR (Depth Image Based Rendering) techniques [1]. The manual 2D-to-3D video conversion workflow basically consists of three steps: rotoscoping/segmentation, depth assignment and inpainting (hole filling) [2]. Manual or semi-automatic methods are often very labour intensive but lead to far better results than the automatic ones. The creation of "fake" 3D videos which are often also called post-production 3D (post 3D) videos, has increased with significant rates in the movie and entertainment industry. However, since the quality of "fake" 3D material is often inferior to that of "real" 3D, it is essential that the viewers have the opportunity to know if the movie or TV programme which they watch, is fake 3D or not, as a quality indicator.

In this paper, we propose a novel algorithm for fake 3D video recognition. The main idea of this algorithm lies in the quality assessment of the two views/channels (left-right) through the computation of sharpness no reference metrics on the left and the right frame's pixels in a stripe around a foreground object, in our case a human figure. Distinction between fake and real 3D is based on the fact that the sharpness in this "detection" stripe of the two views differs in fake 3D videos due to the hole filling (inpainting) [3]. Inpainting is an essential step of the 2D-to-3D video conversion methods [4] that follows the creation of the right frame (rendered view) from the left frame (source view) through the horizontal displacement of certain image elements [5], in order to create disparity. This displacement leads to the creation of holes (image regions with no color/texture information) in the rendered view. Inpainting of the holes created in the rendered view due to object displacement is usually a low-pass procedure. As a result, the sharpness of the inpainted areas around objects in the rendered view is different from the sharpness of the same areas in the source view in fake 3D videos. In contrast, the sharpness in such areas in the left and right view of real 3D videos is usually the same. This fact is used by our algorithm in order to distinguish between real and fake 3D videos.

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287674 (3DTVS). The publication reflects only the authors views. The EU is not liable for any use that may be made of the information contained herein.

As far as we know, no method of fake 3D video recognition has been proposed so far in the literature. The only somehow related paper is [6] that presents a method for the automatic detection of edge-sharpness mismatch in converted stereoscopic 3D (fake 3D) videos. In such videos, an object (e.g. human figure) boundary is presented sharper in one view and blurrier in the other, yielding binocular rivalry. To detect this problem, the authors of this paper estimate the disparity map, extract boundaries of the object with the Canny edge detector and analyze edge-sharpness correspondence between the two views with sharpness estimation metrics that rely on color and texture information.

This paper is organized as follows: Section 2 provides the details of the proposed algorithm of fake 3D video recognition. In Section 3 we present the dataset and the experiments which have been conducted to measure the method performance. Finally, conclusions are presented in Section 4.

2. PROPOSED METHOD

The proposed algorithm for fake 3D video recognition with the use of sharpness estimation metrics applied on stripes around foreground human figures in the left and the right frame of 3D video consists of the following steps:

1. Stereo Frame Pair Selection: The frames on which the algorithm is applied, are manually selected (Figure 1a,b). More specifically, we select a few frames which present a human figure in a medium close up view. However, the frame selection can be automated by applying a face detection algorithm to the video and selecting frames in which the faces are large enough.

2. Disparity Map Estimation: Extraction of the disparity map of the left view of the selected stereo frame pair (Figure 1c).

3. Foreground Object Segmentation and Binary Mask Creation: The disparity map is segmented in homogenous regions via the graph based image segmentation algorithm proposed in [7] (Figure 1d). Then, a region that corresponds to a human figure is selected. This selection is currently done manually. However, it can be automated by calculating the overlap of the facial bounding box generated by a face detector with the regions resulting from the segmentation. The region which has the largest overlap can then be selected as the one representing the human figure. Subsequently, a binary mask representing the area covered by the human figure is created for the left channel/view (Figure 2a). Such a mask is also created for the right channel (Figure 2b) by horizontally shifting the left channel mask. The horizontal displacement is evaluated by finding the position of the left frame mask on the right frame where the mean square error between the image content of the left frame beneath the mask and the image content of the right frame beneath the horizontally transposed mask is minimum.

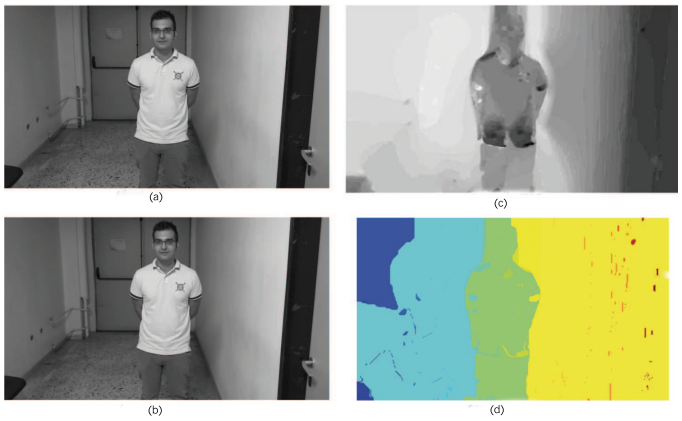

Fig. 1: (a) left view, (b) right view, (c) disparity map, (d) segmentation results

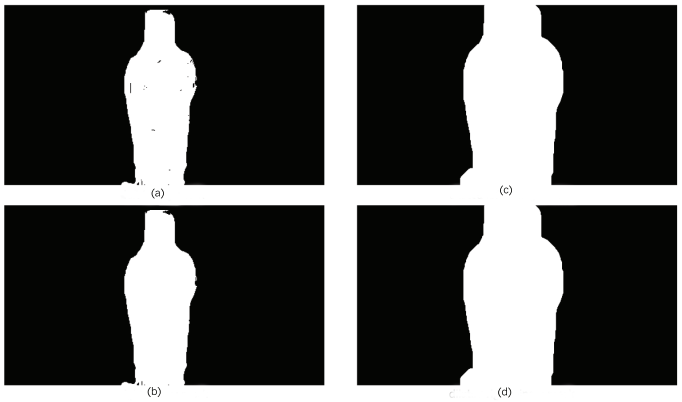

Fig. 2: (a) left binary mask, (b) right binary mask, (c) dilated left binary mask, (d) dilated right binary mask

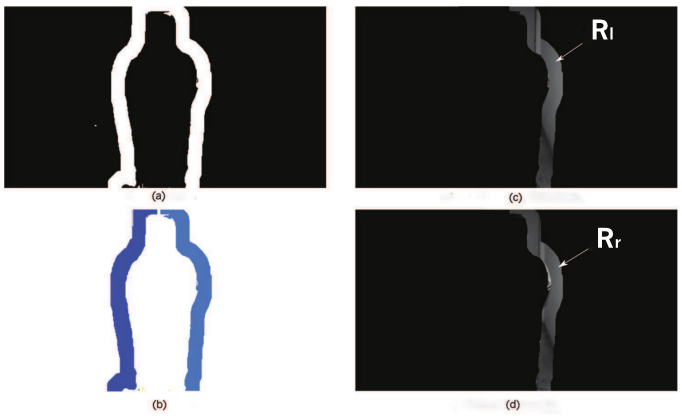

Fig. 3: (a) stripe creation, (b) left or right part of stripe selection, (c) stripe selection on the left frame, (d) stripe selection on the right frame

4. Creation of the Detection Stripe: Since the algorithm is based on the sharpness difference in areas around foreground

objects such as human figures, a so called detection stripe is created around the left and right binary masks (Figure 3a). This is done by first applying binary dilation to the two masks (Figure 2c,d). The dilation is applied with a disk structuring element whose size is equal to the mean disparity of the segmented region. If the mean disparity of the segmented region is small then the size of the structuring element is selected to be equal to the mean disparity of the bounding box of the human figure outside this figure. This is because in this case, it is the background that has been shifted rather than the human figure. The detection stripes are then created by subtracting respectively the left binary mask from the dilated left binary mask and the right binary mask from the dilated right binary mask (Figure 3a).

5. Selection of a Part of the Detection Stripe: As already mentioned, the shifting of the human figure towards the left or the right in the right channel (rendered view) during the 2D to 3D conversion procedure creates a hole (namely a stripe without image content whose width is approximately equal to the amount of shifting of the figure) to the right or to the left of this figure in this channel. This hole, which is subsequently inpainted by the 2D-to-3D conversion algorithm, is most probably included in the detection stripe due to the way the width of this stripe is evaluated (see previous step) and is the target of analysis of the proposed algorithm. Thus, the part of the detection stripe where the hole might be present is selected by checking the direction of translation of the binary mask (see step 3). If the mask was translated to the right (left), then the human figure was translated in the same direction and thus the hole is on the left (right) part of the right frame detection stripe. Thus, this part of the stripe is selected. This is done by finding the topmost point of the stripe, splitting the stripe into two parts by using the vertical line that passes from this point and keeping the appropriate half of it. This procedure is executed in both stripes, on the left and the right frames (Figure 3b,c,d) resulting in two regions R_l, R_r, one per frame. It should be noted that if during step 4, a left (right) shift of the background (rather than of the figure) has been detected, the left (right) part of the detection stripe is retained.

6. Sharpness Estimation: A sharpness estimation metric is computed on the left/right frame pixels which are contained in regions R_l, R_r i.e., in the selected parts of the detection stripes around the human figure (Figure 3c,d). Two different metrics are used: the Cumulative Probability of Blur Detection (CPBD) metric that is based on the image gradients [8,9] and the S_3 metric that is based on the slope of the spectrum magnitude and the total spatial variation [10]. Both metrics are scalar, belong to the class of no reference metrics (i.e., they require no reference image for their evaluation) and obtain a low value when the image is not sharp (as in areas filled by inpainting) and a high value otherwise. In addition to providing a single scalar value, the method used for the calculation of the S_3 metric generates also an S_3 sharpness map for the image or image region under consideration. The pixel values in this map corespond to the estimated perceived local sharpness.

7. Frame and Movie Level Classification: Each selected frame is classified as fake 3D or real 3D. This is accomplished using two different approaches:

a) Classification of the frames using Support Vector Machines (SVM): First, a feature vector is computed for every selected stereo frame. The feature vector, which consists of N elements is calculated as follows: the N pixels $[P^l_{(1)}...P^l_{(N)}]^T$ of the S_3 map of region R_l with the largest values are found. Subsequently, the differences of the pixels' values from the values of the correspoding pixels of the S_3 map of region R_r are calculated and used to form the feature vector **f**:

$$\mathbf{f} = [P^l_{(1)}...P^l_{(N)}]^T - [P^r_{(1)}...P^r_{(N)}]^T \quad (1)$$

The two-class SVM classifier is then trained with the feature vectors of frames in videos beloging to the training set and subsequently applied on frames of videos from the test set in order to classify them as "fake" or "real" 3D.

b) Classification of frames by thresholding: According to this approach, the absolute difference d between the S_3 or CPBD scalar values for regions R_l and R_r is evaluated:

$$d = \left| S^{R_l} - S^{R_r} \right| \quad (2)$$

where S is either the S_3 or the CPBD value. d is then compared to an appropriate threshold T. Small values of d, i.e. values $0 < d < T$ denote that the sharpness in the selected parts of the detection stripe in the right and left frame (regions R_l and R_r) is similar and the frame is characterized as real 3D. On the other hand, values larger than the threshold i.e. $0 < T < d$ lead to the frame being classified as fake 3D. Indeed, large values of d denote that the selected parts of the stripes in the two frames have significantly different sharpness, a phenomenon that is most probably the result of inpainting in the rendered (right) view, that leads to a much more smooth region compared to the source view. Selection of the threshold value T is done by a cross validation approach that is detailed in Section 3.

Fake/non fake decision at the level of an entire video or movie, which is the final aim of the algorithm, is taken using a majority rule. More specifically, when the algorithm is applied on several frames from a movie, each is being characterized as fake 3D or real 3D. The entire movie is then characterized according to the characterization assigned to the majority of the frames.

The mean computational time needed in order to reach a decision for a frame using the proposed method is 55.22 seconds for the SVM based variant and 54.05 seconds for the threshold based variant. These times refer to a computer with Intel Core 2 Duo 2 GHz processor and 4GB RAM.

3. EXPERIMENTAL RESULTS

The efficiency of the fake 3D video recognition algorithm was examined on a dataset consisting of sample stereo frame pairs from stereo movies (fake and real 3D ones). In more detail, the algorithm was applied on stereo frame pairs from 28 feature length movies: 14 filmed in 3D and 14 converted in 3D in post-production. Information for the category (fake/real 3D) each movie belongs to was obtained from [11]. The movies are in High Definition (HD) (1920×1080). Five stereo frame pairs from every stereoscopic movie have been used for the application of the algorithm and a majority rule was used to reach a decision for the entire movie as described in Section 2.

Both variants of the algorithm namely the one using Support Vector Machines and the one that utilizes thresholding were evaluated. Moreover, within the second variant both the S_3 and the CPBD metrics were utilized.

3.1. Classification with Support Vector Machine (SVM)

The 2 class-SVM is trained on the 27 movies (135 frames) and it is tested on the remaining movie (5 frames), thus conducting a 28-fold cross validation. Various values of N (feature vector dimension) and different SVM kernels (linear, polynomial etc) were used and the best results are reported here. The best accuracy of SVM classification at the movie level is 85.71% at N=400 dimensions of feature vector with polynomial kernel.

3.2. Classification by thresholding

A 28-fold cross validation was also used for the experimental evaluation of the threshold based variant of the proposed method. In more detail, in each one of the 28 folds, 27 movies from the dataset are used to evaluate (through range search) the threshold value T that leads to the best classification accuracy and this value is used to classify the remaining movie. The classification accuracy for the CPBD and S_3 sharpness metrics was 78.57% and 71.43% respectively. It is obvious that both metrics lead to good results, with the CPBD metric providing the best results which are however somewhat inferior to those obtained by the SVM-based approach.

4. CONCLUSIONS

In this paper, we have presented a method for distinguishing fake 3D from real 3D movies that is based on the fact that the sharpness around a foreground object (e.g. human figure) differs in the left and the right view of the fake 3D stereo frame pair because of the inpainting (hole filling) procedure.

The two variants of the proposed method have been tested on 28 feature length movies with very good results.

Future work includes automating steps 1 and 3 of the algorithm and testing it to larger datasets.

REFERENCES

[1] Liang Zhang, Carlos Vazquez, and Sebastian Knorr, "3D-TV Content Creation: Automatic 2D-to-3D Video Conversion," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 372 – 383, March 2011.

[2] Aljoscha Smolic, Peter Kauff, Sebastian Knorr, Alexander Hornung, Matthias Kunter, Marcus Müller, and Manuel Lang, "Three-dimensional video postproduction and processing," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 607–625, 2011.

[3] Ming Xi, Liang-Hao Wang, Qing-Qing Yang, Dong-Xiao Li, and Ming Zhang, "Depth-image-based rendering with spatial and temporal texture synthesis for 3DTV," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–18, February 2013.

[4] Liang Zhang and Wa James Tam, "Stereoscopic Image Generation Based on Depth Images for 3DTV," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 191 – 199, June 2005.

[5] Benoit Michel, *Digital Stereoscopy: Scene to Screen 3D production workflow*, Stereoscopy News, 2013.

[6] Alexander Bokov, Dmitriy Vatolin, Anton Zachesov, Alexander Belous, and Mikhail Erofeev, "Automatic detection of artifacts in converted S3D video," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, March 2014, pp. 901112–901112.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167 – 181, September 2004.

[8] Narvekar Niranjan and Karam Linti, "A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678 – 2683, March 2011.

[9] "CPBD Sharpness Metric Software," http://ivulab.asu.edu/Quality/CPBD, March 2011.

[10] Cuong T Vu, Thien D Phan, and Damon M Chandler, "S3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 934–945, March 2012.

[11] "Is it Real or Fake 3D?," http://www.realorfake3d.com/, Accessed: 2015-02-09.