IMMERSIVE 3D HOLOSCOPIC VIDEO SYSTEM

Amar Aggoun[1], Emmanuel Tsekleves[1], Dimitrios Zarpalas[2], Anastasios Dimou[2], Petros Daras[2], Paulo Nunes[3,4], Luís Ducla Soares[3,4]

[1]Brunel Universty (UK),

[2]Informatics and Telematics Institute (Greece)

[3]Instituto de Telecomunicações, [4]Instituto Universitário de Lisboa (ISCTE-IUL) (Portugal)

ABSTRACT

3D holoscopic imaging is employed and presented here as part of a three-dimensional imaging system, allowing the display of full colour images with continuous parallax within a wide viewing zone. In this paper, recent advances made by the authors with respect to the 3D holoscopic imaging technology from the point of view of optical systems and 3D image processing including 3D image coding, depth map computation and search and retrieval are discussed.

Index Terms – 3D video, holoscopic video, glassless 3D video

INTRODUCTION

Creating 3D content has been the goal of many researchers in the academia and industry as well as artists for many years (e.g., in cinema, television and performing arts). Recent film releases such as 'Avatar' have revolutionalised cinema by the extensive use of 3D technology and 3D content production along with real actors creating a new genre at the outset of the 2010s. The success of 3D cinema has led several major consumer electronics manufacturers and broadcasters to launch 3D-capable Televisions (TVs) and offer 3D content. Today's 3DTV technology is based on stereo vision, where left and right eye images are presented to the viewer through temporal or spatial multiplexing by wearing a pair of glasses. The next step in the 3DTV development is expected to be the multiview autostereoscopic imaging system, where a large number of pairs of video signals are recorded and presented on a display that does not require glasses for viewing [1]-[2]. Although, several autostereoscopic displays have been reported, there are still limitations on resolution and viewing position. Furthermore, stereo and multiview technologies rely upon the brain to fuse the two disparate

images to create the 3D sensation. As a result, such systems tend to cause eye strain, fatigue and headaches after prolonged viewing as users are required to focus on the screen plane (accommodation) but to converge their eyes to a point in space in a different plane (convergence), producing unnatural viewing. With recent advances in digital technology, some of these human factors, which result in eye fatigue, have been eliminated. However, some intrinsic eye fatigue factors will always exist in stereoscopic 3D technology [3].

The above facts have motivated researchers to seek alternative means for capturing true 3D content, two of the most recognised being holography and holoscopic Imaging. Due to the interfering of coherent light fields required to record holograms, their use is still limited and mostly confined to research laboratories. Holoscopic imaging (also referred to as Integral Imaging) in its simplest form consists of a lens array mated to a digital sensor with each lens capturing perspective views of the scene [4]-[10]. The light field in this case does not need to be coherent and 'holoscopic' colour images can be obtained with full parallax. This conveniently allows more conventional live capture and display procedures to be adopted. Furthermore, 3D holoscopic imaging offers fatigue free viewing to more than one person, independently of the viewer's position. With recent advances in the theory and microlens manufacturing, 3D holoscopic imaging is becoming a practical and prospective 3D display technology and is attracting much interest in the 3D area. It is now accepted as a strong candidate for next generation 3D TV [3]. A project funded by the EU-FP7 ICT-4-1.5 – Networked Media and 3D Internet, entitled "3D Live Immerse Video-Audio Interactive Multimedia" (3D VIVANT) offers a number of advances in the 3D holoscopic imaging technology for capture, representation, processing and display of 3D holoscopic content [4].

In this paper, an end-to-end 3D holoscopic video framework is described (from capturing, to processing and visualizing); overcoming most of the aforementioned restrictions the current 3D technologies suffer from.

3D HOLOSCOPIC CONTENT GENERATION

3D holoscopic imaging is a technique that is capable of creating and representing a true volume spatial optical model of the object scene in the form of a planar intensity distribution, by using unique

optical components. A 3D holoscopic image is recorded using a regularly spaced array of small lenslets closely packed together in contact with a recording device as shown in Figure 1a [5]. Each lenslet views the scene at a slightly different angle to its neighbour and therefore a scene is captured from many view points and parallax information is recorded. The replay of the 3D holoscopic images is achieved by placing a microlens array on the top of the recoded planar intensity distributions that is illuminated by diffuse white light from the rear. The object will be constructed in space by the intersection of ray bundles emanating from each of the lenslets as shown in Figure 1b. In replay, the reconstructed image is pseudoscopic (inverted in depth). Optical and digital techniques to convert the pseudoscopic image to an orthoscopic image have been proposed [6]-[10] in the last two decades.

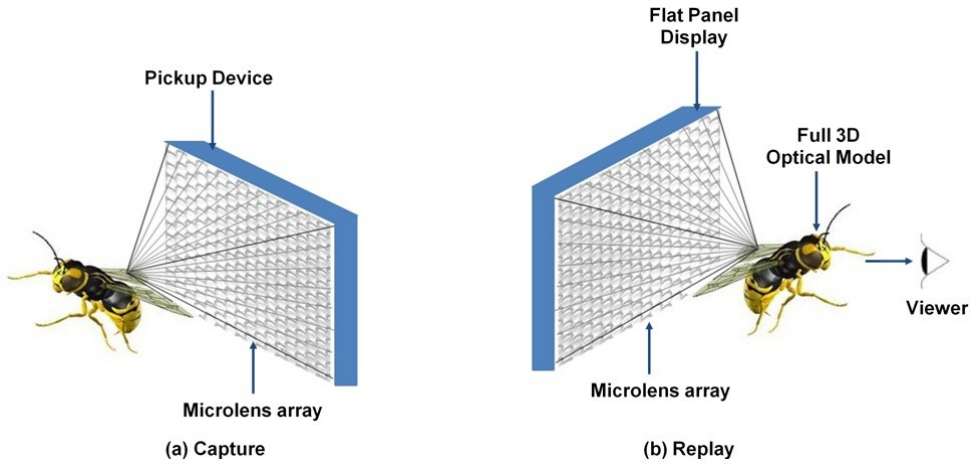


Figure 1: Recording and replay of a 3D Holoscopic Imaging System.

A disadvantage of the camera setup in Figure 1a is that it does not possess depth control; this means that the reconstructed object appears in its original location in space and therefore can only produce 3D virtual images or 3D real pseudocsopic images. Furthermore, objects very far from a microlens array will suffer from poor spatial sampling on sensor pixels; therefore this type of camera would be well suited to close imaging applications. In addition standard methods of microlens manufacturing e.g. UV and hot embossing have shrinkage and replication errors which can yield errors in the pitch of ±20µm/20mm (35µm/35mm), which is particularly a large percentage error for those microlens array's with small pitch.

To remedy the two problems an objective lens and a relay lens are added as shown in Figure 2.

An objective lens is added to provide depth control, which allows the image plane to be near the microlens array, the spatial sampling of the 3D holoscopic image is given by the number of lenses and thus higher resolution images can be obtained by reducing the size of the lenses. There is the tradeoff between the number of lenses and number of viewpoint images or pixels under each lens. These pixels define discrete locations in the aperture of the objective lens from where the object can be viewed. Therefore making the lenses smaller reduces the angular information about points in the object as fewer pixels can view it.

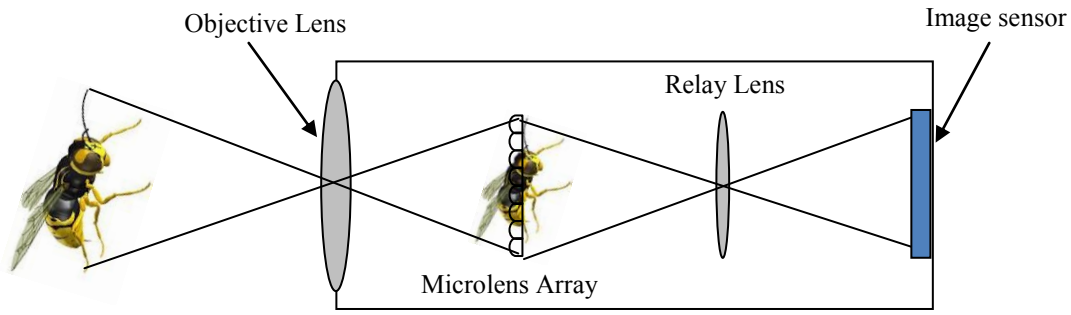
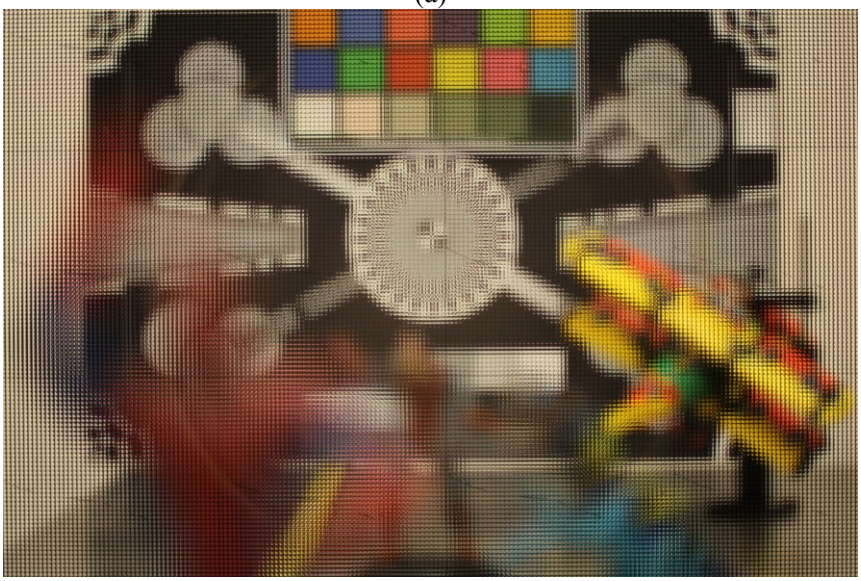


Figure 2: 3D holoscopic camera with a relay and objective lens.

Live images are recorded in a regular block pixel pattern. The planar intensity distribution representing a 3D holoscopic image is comprised of 2D array of M×M micro-images due to the structure of the microlens array used in the capture. A significant problem with the setup shown in Figure 2 is the viginetting of micro-images at the microlens pickup stage due to the use of an image relay. This is resolved by introducing a field lens at the back of the microlens array. Figure 3a and 3b show images obtained without and with the field lens. The purpose of a field lens is to relay the image of the pupil in order to prevent vignetting of off axis rays when employing a system with an intermediate image plane. The field lens focal length is determined by the exit pupil and relay entrance pupil conjugates as measured from the microlens array image plane. As it can be seen from Figure 3b most of the vigneting has been removed by the use of the field lens.

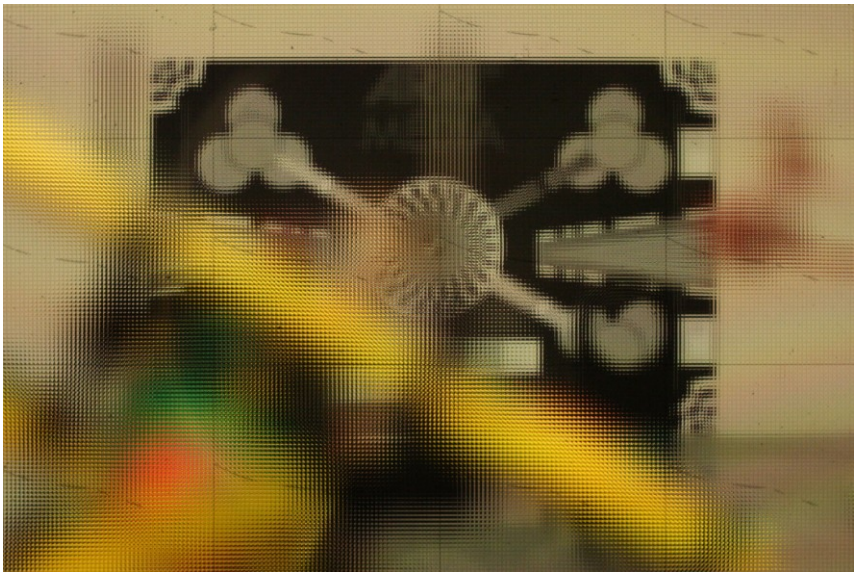
(a)



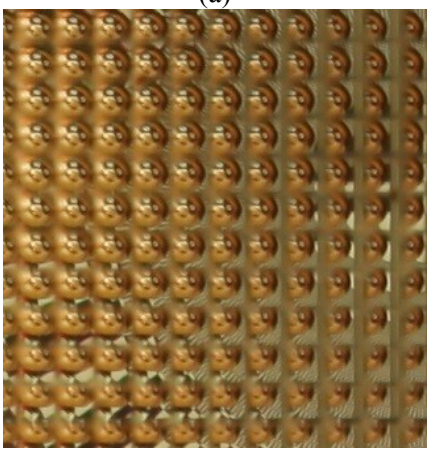
(b)

Figure 3: Holoscopic images captured a) without a field lens b) with a field lens

To achieve above 95% fill factor, a square aperture is fitted on the front of the camera. The square aperture at the front of the camera and the regular structure of the square microlenses array used in the square grid (recording microlens array) gives rise to a regular structure in the intensity distribution as illustrated in Figure 4.



(a)



(b)

Figure 4: (a) Recorded 3D Holoscopic Image using a 90μm pitch microlens array. (b) Magnified section.

3D HOLOSCOPIC CONTENT VISUALISATION

One of the aims of 3D VIVANT project is to investigate the possibility to display 3D holoscopic content on commercially available autostereoscopic displays. Currently there are a number of autostereoscopic displays available on the market which uses the combination of lenticular optics in combination with LCD panels. In current autostereoscopic displays, the lenticular elements are positioned at an angle to the LCD pixel array [11]. This mixes adjacent views reducing image flipping problems and spreading the effect of the black mask making it less visible. The other benefit of this design is that each view has a better aspect ratio, rather than splitting the display horizontally into many views both horizontal and vertical directions are split. The slanted lenticular arrangement will require sub-pixel mapping to allow all pixel along a slanted line to be imaged in the same direction.

To adapt a 3D holoscopic content to be viewed on an autostereoscopic display, unidirectional 3D holoscopic images are used. These images are obtained using a special case of the 3D holoscopic imaging system where 1D cylindrical microlens array is used for capture and replay instead of a 2D array of microlenses. The resulting images contain parallax in the horizontal direction only and can be displayed using a lenticular sheet in association with an LCD panel. A 3D holoscopic virtual camera is used to generate eight orthographic views and a software tools is developed which performs the sub-pixel mapping and interlacing of the orthographic views into a single image to be fed into the 3D autostereoscopic display. Figure 5 shows a photograph of a real hand in front of the screen showing a computer generated 3D holoscopic image.

Figure 5: Two perspective views of a 3D holoscopic image on a commercial 3D autostereoscopic display. Parts of the 3D holoscopic image on the screen (such as the 1st Globe) can be seen to be at the same depth as the image of the hand.

CODING AND TRANSMISSION

In order to make 3D holoscopic video delivery feasible over real-world networks and storage media, video codecs will have to deal efficiently with the large amount of data captured with such high resolution sensors. Consequently, new efficient video coding tools become of paramount importance.

Due to the small angular disparity between adjacent microlenses, a significant cross-correlation exists between neighboring micro-images. Therefore, this inherent cross-correlation of 3D holoscopic images can be seen as a type of self-similarity and may be exploited for improving coding efficiency. For example, Figure 6 shows two 3D holoscopic video frames with a different amount of self-similarity redundancy.

A scheme for self-similarity estimation and compensation was proposed in [13], in order to explore the high self-similarity between neighboring micro-images and improve the performance of H.264/AVC. The self-similarity estimation process uses a block-based matching in an area of the current picture that has been already coded and reconstructed, to find a best match, in terms of a suitable matching criterion (e.g., the sum of absolute differences), for prediction of a block area. This prediction process has the advantage of being able to find efficient predictions for 3D holoscopic content without the need to know the precise structure of the underlying microlens array, and consequently, the arrangement of the micro-images.

Recently, a new standardization project on video coding, called High Efficient Video Coding (HEVC) [12], is coming up by joint efforts from MPEG and ITU-T that formed the Joint Collaborative Team on Video Coding (JCT-VC). This new generation video coding standard promises to improve the coding efficiency of the state-of-art H.264/AVC to fulfill the current compression efficiency requirements, among other requirements, for high and ultra-high resolution video content. However, HEVC does not currently efficiently handle 3D holoscopic content, since it does not take into account its inherent self-similarity. This can be overcome by combining the flexible coding tools from HEVC with the self-similarity estimation concept [13][14], by adding new predictive coding modes, which exploit the special arrangement of 3D holoscopic content. In [14], this is done by defining two new prediction modes for HEVC referred to as Self-similarity (SS) and SS-skip prediction modes, which take full advantage of the inherent structure of 3D holoscopic images.

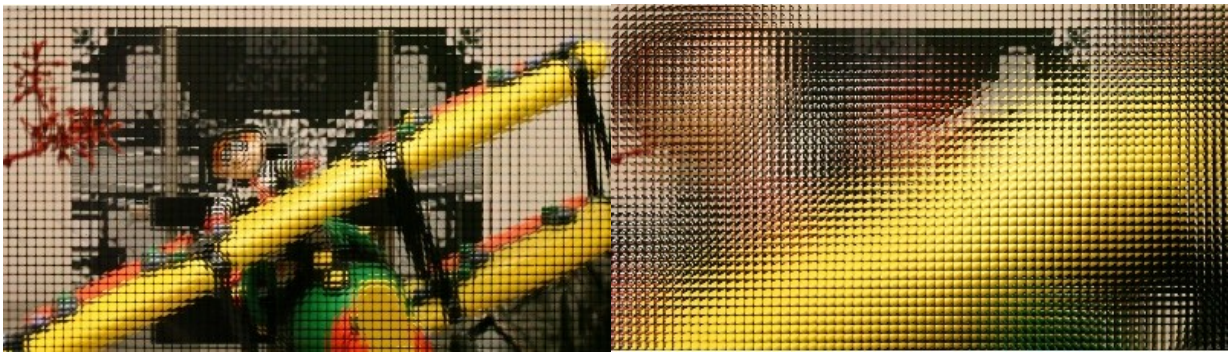
The Self-similarity prediction mode, which is based on the HEVC Inter prediction mode, enables all the original prediction unit (PU) partition patterns, but replaces the motion estimation and compensation with the self-similarity estimation and compensation process. When this mode is used, the relative position between a given PU and its prediction is encoded and transmitted as a vector (similarly to a motion vector). This vector is referred to as the self-similarity vector [13][14]. In addition to this, the prediction residual is also encoded and transmitted.

In self-similarity estimation, the search range should be adapted to the size of the PU being considered. In the self-similarity compensation process, for each PU, one self-similarity vector is

derived. In addition, an adapted scheme based on the advanced motion vector prediction (AVMP) is used to select a self-similarity vector prediction.

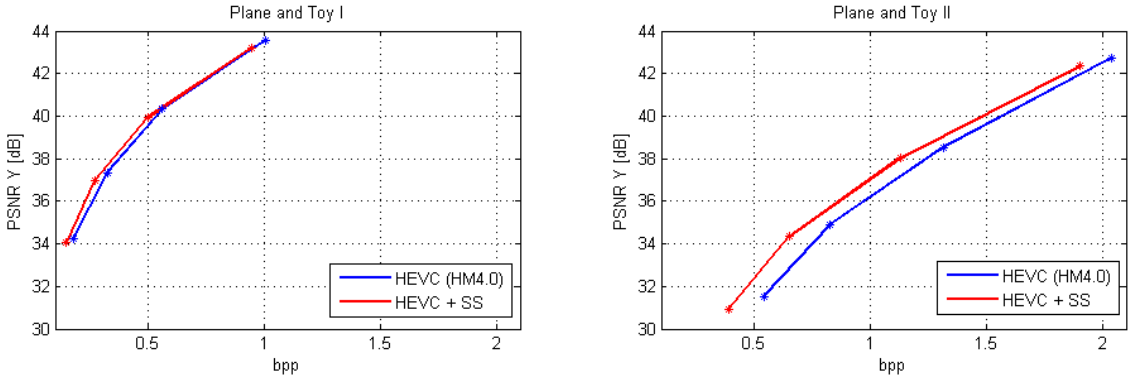
On the other hand, when the SS-skip prediction mode is used, only a self-similarity vector is encoded and transmitted for each PU (i.e., no residual data is sent to decoder). This mode infers a self-similarity vector for the current PU by considering self-similarity vector candidates from spatially neighboring PUs. These inferred self-similarity vectors are constrained to point to the area defined in the self-similarity estimation process since this guarantees that the prediction indicated by the chosen self-similarity vector (which belongs to the same frame) is already available as a self-similarity reference.

Applying this coding scheme to 3D holoscopic content, such as the *Plane and Toy* sequence illustrated in Figure 6 shows that the HEVC performance can be significantly improved for this type of content (see Figure 7). Moreover, it is important to notice that the improvements are highly dependent on the amount of self-similarity redundancy existing in the video frames, as can be seen in Figure 7a and Figure 7b. As such, it is possible to conclude that the stronger the self-similarity redundancy is, the better the performance of the self-similarity coding approach.



(a) (b)

Figure 6: 3D holoscopic test sequence *Plane and Toy* captured using a 250μm pitch microlens array: (a) frame with less self-similarity redundancy; (b) frame with more self-similarity redundancy



(a) (b)

Figure 7: PSNR results for the *Plane and Toy* test sequence: (a) frame with less self-similarity redundancy; (b) frame with more self-similarity redundancy

CONTENT INTERACTIVITY: DEPTH EXTRACTION, SEGMENTATION AND RETRIEVAL

3D holoscopic video allows the user to see different perspectives of the same 3D scene by simply changing his/her position. However, in order for the viewing experience to be a truly interactive experience, the user should be able to individually select the 3D objects in the 3D holoscopic video scene and hyperlink to other content by clicking on them and search and retrieve the desired 3D content. This paper discusses methods 3D holoscopic object extraction and segmentation as well as search and retrieval.

Anchoring Graph Cuts towards Accurate Depth Estimation in Holoscopic Images

The compactness of using 3D holoscopic imaging in depth measurement has been attracting attention as novel depth extraction techniques are reported [15]. In [16], a method for depth computation is presented, that involves the extraction and matching a set of feature points, called Anchor Points, and then trying to fit a surface to the reconstructed features that aims to optimally connect them. The method was based on its two ancestor techniques of [17], and after thorough analysis of the advantages and disadvantages of those two methods, the new framework was derived. As such, it proposes one surface fitting optimization that is implemented through graph cuts, which is constrained by the anchor points. Anchor points are first extracted, which are the 3D points of the strong correspondences among a large number of successive viewpoint images. In every set of seven

successive viewpoint images, one of them is regarded as the central one, and is matched with the rest. In case the stereo pairs of the same pixel from the central viewpoint, with the majority of the rest images, agree on the same depth value, then this estimated 3D point is regarded as a very reliable one and is considered as an Anchor Point.

Anchor points serve on the next step for constraining the optimization procedure, which otherwise can easily get stuck to local extrema, due to the high complexity of the optimization. Anchoring the optimization results in enhanced estimation accuracy along with reduction in the optimization complexity. In the proposed formulation, graph cuts are trying to find the optimized solution of the depth of the 3-D points that correspond to each pixel, instead of the disparities between pixels of adjacent viewpoint images. The 3-D scene is scanned uniformly along the emanating projection rays of the integral image. Furthermore, the proposed framework enables the modeling of the piecewise nature of depth by introducing a twofold regularization term among adjacent pixels on both the elemental and viewpoint images. This twofold neighborhood handling leads to reconstructed scenes with high spatial smoothness. The proposed algorithm's workflow is depicted in Figure 8.

The results of the proposed algorithm for synthetic 3D holoscopic images are illustrated in Figure 9, where the estimated depth-map is shown in contrast with the scene's actual depth-map, both being estimated from the central's viewpoint angle. In all cases, the different objects are correctly perceived and differentiated from their neighbouring, based on their estimated depth values. The reliability of the depth estimations of the Anchor Points is verified by their mean relative error, which is about 3,98% in the synthetic dataset[1]. The mean relative error for the complete image depth estimates is 6,13% for the whole synthetic dataset. Results for real 3D holoscopic images are shown in Figure 10.

[1] ftp://ftp.iti.gr/pub/Holoscopy/

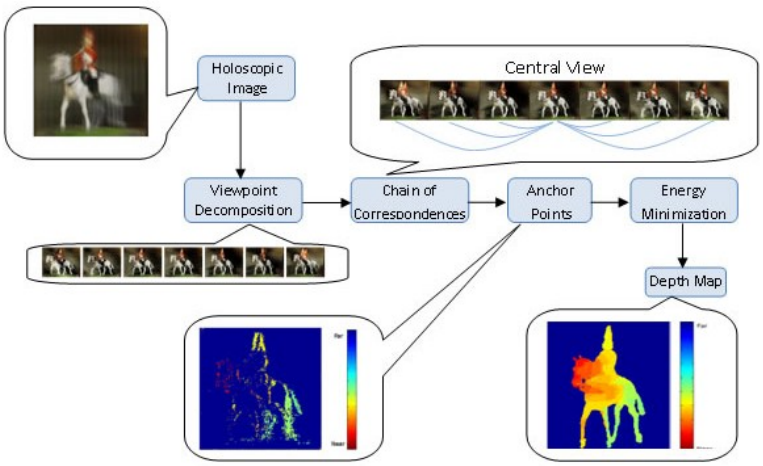


Figure 8: Schematic representation of the proposed procedure for depth estimation.

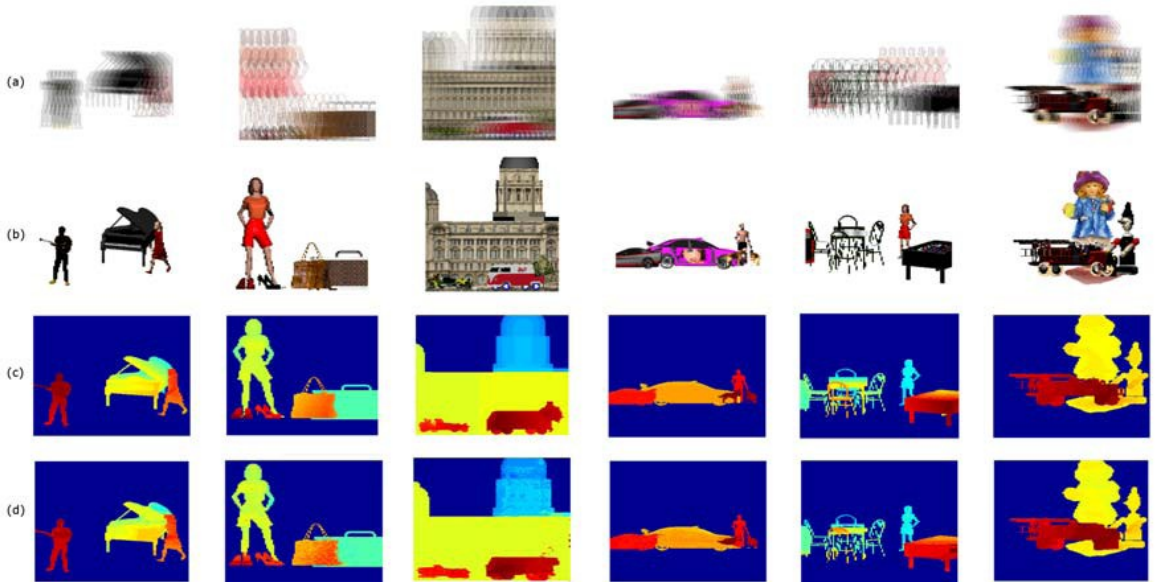


Figure 9: Results of the proposed algorithm on synthetic holoscopic images. (a) Holoscopic images. (b) Corresponding viewpoint images. (c) Actual depth map of the scene. (d) Depth map estimation using the proposed method.

Search and Retrieval of 3D holoscopic images

To be able to effectively browse into collections of 3D holoscopic images, a search and retrieval framework has been created by exploiting the rich information contained in them. Due to its nature, each holoscopic image is analysed into three distinct visual modalities: the 2D viewpoint images, the depth information either as a depth-map or as a 2.5D point cloud, and the 3D curvature information extracted from the depth-map, again as a 2D curvature map.

For each one of the modalities, different types of state-of-the-art descriptors are extracted, in

order to choose the optimal combination. These descriptors are applied on the viewpoint images, the depth-maps and the curvature maps. Furthermore, on the 2.5D point cloud depth modality, a novel local shape descriptor for 2.5D images, called the Projection Images [18], is also used. For a given point, the projection distances of the projected neighboring points are captured, and are separated into two sets, i.e. the positive projection image and the negative, based on which side of the projection plane they lay. The plane is divided into sectors by defining angular and radial divisions and the mean values of the positive and the negative projection distances for each sector are calculated. This way, a very compact and descriptive twofold representation of the object's shape is captured.

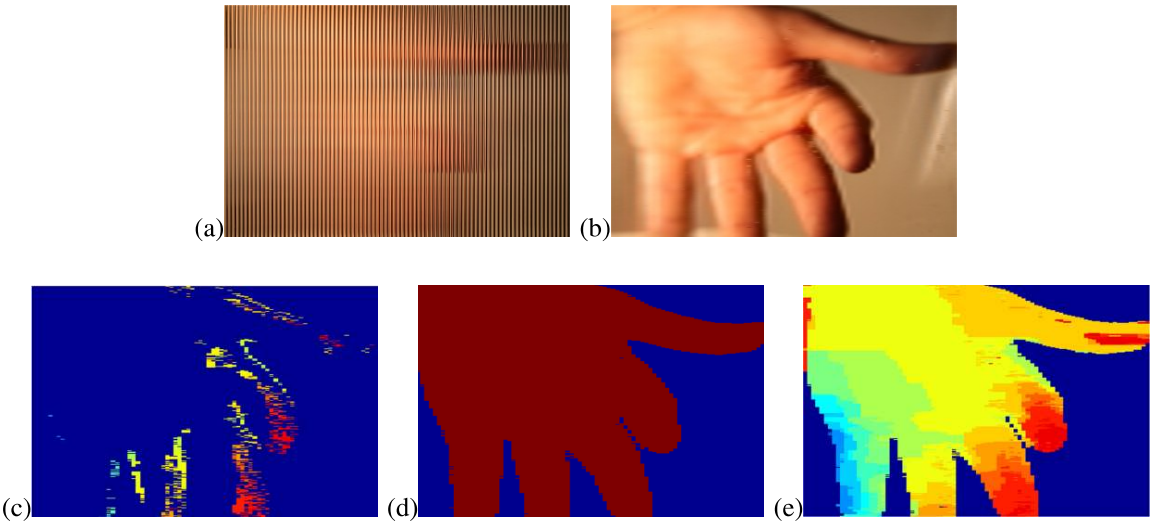


Figure 10: Results on two real holoscopic images captured with cylindrical lenses, depicting a palm and a horseman. For the palm, the holoscopic image is depicted (a) and in (b) one viewpoint image, while (c) depicts the anchor points, (d) the foreground mask, and (e) the extracted depth-map using the proposed method.

A large number of local descriptors are computed for each holoscopic image per modality and the bag-of-words methodology is utilized to create a single descriptor vector for every image. A number of randomly selected local descriptor vectors are used to create clusters that are described as "visual words", using the k-means algorithm. The histogram depicting the appearance frequency of those visual words in each image constitutes the final descriptor vector.

To overcome the lack of spatial information associated with local descriptors, the spatial pyramidal decomposition is used [19]. A wide range of distance metrics is also tested to identify the most suitable

one for each descriptor type. Then, for each modality, the different descriptors are combined through weighting each of them, by optimizing, using the Graph-Cuts technique, the retrieval performance on a training dataset.

Concluding on the combined descriptor vectors for each modality, the next step involves the fusion of the different modalities into a multimodal framework by combining all mono-modal distance matrices to produce an over-all-modalities distance matrix. The latter is achieved by using a Manifold Learning approach which relies on Laplacian Eigenmaps utilising a Heat Kernel so as to construct a multimodal low-dimensional feature space. In this space each data point represents a holoscopic image. Such a framework allows linking the holoscopic content with other types of content (such as 2D photographic images, 3D models, 2.5D point clouds, etc).

Experimental Validation

For the experimental validation of the proposed framework, a database containing 665 synthetic images was created, with nine different classes, namely: “Animals”, “Beds”, “Buildings”, “Cars”, “Chairs”, “Couches”, “People”, “Plants”, “Tables”. Each class contains synthetic holoscopic images consisting of their viewpoint images, along with ground truth of depth and curvature images and the estimated ones.

Each one of the modalities was assessed using precision-recall curves and Mean Average Precision (MAP). The corresponding results are depicted in Figure 11 and Table 1.

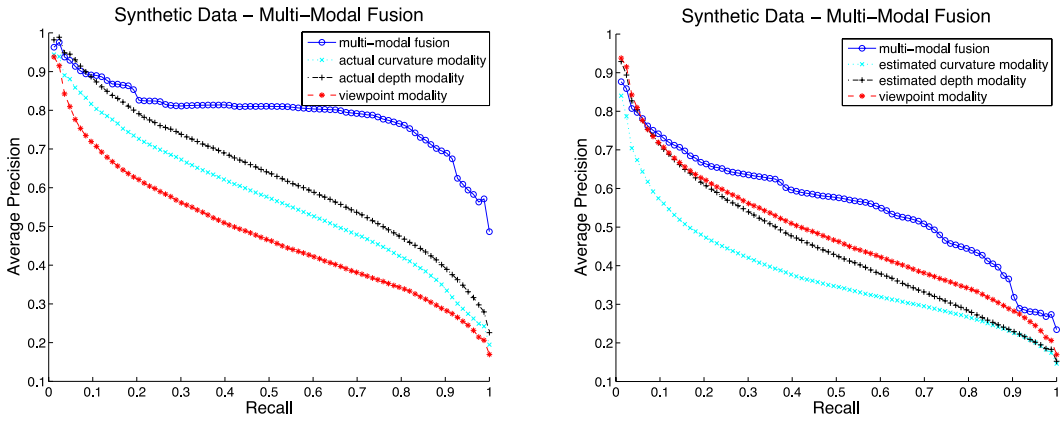


Figure 11: Precision-Recall for all modalities and proposed framework, comparing ground truth (actual) depth (left) vs estimated (right).

Table 1: Mean Average Precision for all modalities and proposed multi-modal fusing framework, using ground truth depth and estimated.

Actual depth		**Estimated depth**	
Modality	**MAP**	**Modality**	**MAP**
mmFusion	0.7936	mmFusion	0.5573
Curvature	0.5722	Curvature	0.3744
Depth	0.6325	Depth	0.4488
Viewpoint	0.4820	Viewpoint	0.4820

Results using the actual depth values show that the depth related modalities are more significant than the texture-based ones. Furthermore, the multi-modal fusion framework strongly boosts the retrieval performance, proving that 3D holoscopic images should be regarded as multi-modal content. Results on the estimated depth show that the retrieval performance depends on the depth estimation quality.

FINAL REMARKS

In this paper, a complete 3D holoscopic video system has been proposed and described. A flexible 3D holoscopic imaging optical system is described which resolve the vigneting problem associated with relay systems. A novel approach to the display of 3D holoscopic images on commercially available autostereoscopic displays is demonstrated. A novel coding of the 3D holoscopic video based a modification of the HEVC scheme. A self similarity module is incorporated to take advantage of the inherent structure of 3D holoscopic images. The results show that the self similarity coding approach performs better than the standard HEVC scheme. Furthermore the paper discusses a depth map computation algorithm and search and retrieval tool created by exploiting the rich information contained in 3D holoscopic images. Results show the performance of the search and retrieval is reliant on the quality of the depth map and that the multi-modal fusion boosts the retrieval performance.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the European Commission under the FP7 project 3D VIVANT (Live Immerse Video-Audio Interactive Multimedia).

REFERENCES

[1] Y. Zhu, T. Zhen, "3D Multi-View Autostereoscopic Display and Its Key Technologie", *Proc. of the Asia-Pacific Conference on Image Processing (APCIP 2009)*, vol. 2, pp. 31-35, Shenzhen, China, (2009).

[2] G. Lawton, "3D Displays without Glasses: Coming to a Screen near You Computer", *IEEE Computer*, vol. 44, no. 1, pp. 17-19, (2011).

[3] L. Onural, "Television in 3-D: What are the Prospects", Proc. IEEE, 95(6), (2007).

[4] http://www.3dvivant.eu.

[5] G. Lippmann, "Epreuves Reversibles Donnant la Sensation du Relief", *Journal de Physique Théorique et Appliquée*, vol. 7, no. 1, pp. 821-825, (1908).

[6] A.Aggoun: "3D Holoscopic Imaging Technology for Real-Time Volume Processing and Display", *High-Quality Visual Experience Signals and Communication Technology*, 2010, IV, 411-428, DOI: 10.1007/978-3-642-12802-8_18, (2010).

[7] J.S. Jang and B. Javidi, "Formation of orthoscopic three dimensional real images in direct pickup one step integral imaging", Optical Engineering, Vol. 42(7), 1869-1870, (2003).

[8] M. Okui, F. Okano, "3D Display Research at NHK", *Workshop on 3D Media, Applications and Devices*, Berlin, Germany, (2009).

[9] M. Martínez-Corral, M., *et. al*. "Formation of real, orthoscopic integral images by smart pixel mapping". Optics Express 13, pp. 9175-9180, (2005).

[10] B. Javidi, *et. al.*, "Orthoscopic, long-focal-depth integral imaging by hybrid method" Proc. of SPIE Vol. 6392, 639203, (2006).

[11] C. van Berkel, D.W. Parker, A.R. Franklin, "Multi-view LCD", Proc SPIE Vol. 2653, (1996).

[12] K. Ugur *et al.*, "High Performance, Low Complexity Video Coding and the Emerging HEVC Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 12, pp. 1688-1697, (2010).

[13] C. Conti, J. Lino, P. Nunes, L. D. Soares, P.L. Correia, "Spatial Prediction Based on Self-Similarity for 3D Holoscopic Image and Video Coding," Proc. IEEE International Conference

on Image Processing (ICIP), Brussels, Belgium, (2011).

[14] C. Conti, P. Nunes, L. D. Soares, “New HEVC Prediction Modes for 3D Holoscopic Video Coding”, Proc. IEEE International Conference on Image Processing (ICIP), Orlando, FL, USA, (2012).

[15] Wu, C., M. McCormick, A. Aggoun, S.Y. Kung, “Depth Map from Unidirectional Integral Images using a Hybrid Disparity Analysis Algorithm” IEEE Journal of Display Technology, 4 (1), 101-108, (2008).

[16] D. Zarpalas, I. Biperis, E. Fotiadou, E. Lyka, P. Daras and M.G. Strintzis, "Depth estimation in integral images by anchoring optimization techniques," in IEEE International Conference on Multimedia & Expo (ICME), Barcelona, (2011).

[17] D. Zarpalas, E. Fotiadou, I. Mpiperis, and P. Daras, "Anchoring-Graph-Cuts towards Accurate Depth Estimation in Integral Images," accepted for publication in IEEE/OSA Transactions on Journal of Display Technology, (2012).

[18] D. Zarpalas, G. Kordelas, P. Daras, “Recognizing 3D Objects in cluttered scenes using projection images”, IEEE International Conference on Image Processing (ICIP 2011), Brussels, Belgium, September, (2011).

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in IEEE Conference on Computer Vision & Pattern Recognition (CVPR), New York, (2006).