IMPROVED SPATIAL PREDICTION FOR 3D HOLOSCOPIC IMAGE AND VIDEO CODING

Caroline Conti[1], João Lino[1], Paulo Nunes[1,2], Luís Ducla Soares[1,2], Paulo Lobato Correia[1,3]

[1]Instituto de Telecomunicações, [2]Instituto Universitário de Lisboa (ISCTE-IUL), [3]Instituto Superior Técnico,
Torre Norte - Piso 10, Av. Rovisco Pais, 1, 1049-001, Lisboa, Portugal
phone: + (351) 218418461, fax: + (351) 218418472, email: {caroline.conti, joao.lino, pjln, lds, plc}@lx.it.pt

ABSTRACT

Holoscopic imaging, also known as integral imaging, is promising to change the market for 3D television since it provides a solution for glassless 3D. This paper starts by making a brief presentation of the general concepts behind holoscopic imaging, with a special emphasis on the spatial correlations that are present in this type of content, which are caused by the micro-lens array used both for acquisition and display. Behind this micro-lens array, many micro-images with a high cross-correlation between them are formed; in these micro-images only one pixel is viewable from a given observation point. This high cross-correlation can be seen as a form of self-similarity within the holoscopic image and it can effectively be exploited for coding. Therefore, in order to explore this self-similarity of holoscopic images, a novel scheme for spatial prediction is proposed in this paper. The proposed scheme can be used for coding both still images and intra-frames in video. Experimental results based on a modified H.264/AVC video codec that can handle 3D holoscopic images and video are presented and clearly show the advantages of using this approach.

1. INTRODUCTION

Enriching the user visual experience through more immersive and realistic sensations is nowadays a major driving force in the development visual technologies. After the introduction of colour and high-definition (HD) imaging, 3D promises to be the next step ahead in this area. The actual momentum in 3D content production is a good example of this evolution.

However, current established methods for acquiring and displaying 3D images, such as those based on light polarization or active-shutter techniques, still cannot provide a truly realistic 3D viewing experience in an ergonomic and cost effective manner exhibiting various drawbacks, such as: i) the viewer needs to wear special glasses to get the depth perception [1]; ii) do not provide motion parallax (i.e., a single scene viewpoint is displayed); and iii) may cause visual discomfort.

To overcome some of these drawbacks, holoscopic imaging, also known as integral imaging, is a promising candidate technology since it allows 3D images to be viewed without any special eyewear, exhibiting continuous motion parallax throughout the viewing zone, presenting a variety of many different scene views, depending on the observers' position [2]-[4], with less visual discomfort. Additionally, holoscopic systems are also suited for multiviewer applications and acquisition and display can be achieved, respectively, with a single aperture camera and conventional flat panel displays, equipped with and overlaid micro-lens array. These advantages and the recent advances in sensor and displaying technologies, led 3D holoscopic imaging technology nowadays to be accepted as a strong candidate for next generation 3DTV [5].

Efficient transmission of 3D holoscopic content over limited bandwidth networks requires adequate coding tools that take advantage of the new inherent spatial and temporal data correlations of this type of content.

Several coding schemes for 3D holoscopic intra-frame coding proposed in the literature are based on the three-dimensional discrete cosine transform (3D-DCT) [3][6]-[8]. A common feature of these schemes is the exploitation of the existing redundancy within the micro-images (i.e., images formed behind each micro-lens), as well as the redundancy between adjacent micro-images, by applying the 3D DCT to a stack of several micro-images. These schemes try to explore the inherent spatial structure of 3D holoscopic images, which can be divided into small non-overlapping areas referred to as micro-images.

In alternative to the DCT, other coding schemes are based on the discrete wavelet transform (DWT) [9][10]. In [9], the 3D holoscopic image is decomposed in various sub-images by extracting one pixel with the same position from each micro-image. Each of these sub-images will then be decomposed using a 2D DWT, resulting in an array of coefficients corresponding to several frequency bands. The lower frequency bands of the sub-images are assembled and compressed using a 3D DCT followed by Huffman coding, while the remaining higher bands are simply quantized and arithmetic encoded.

In [11][12], the authors propose to decompose the 3D holoscopic video sequence into multiple sub-image video sequences and to jointly exploit motion (temporal prediction) and disparity between adjacent sub-images to perform compression. In these schemes, the spatial redundancy is exploited by the disparity estimation part of the scheme, similarly to what is done in multiview video coding (MVC) [13]; as such, a precise knowledge of the image structure is needed, notably the sub-image dimensions.

This paper proposes a novel scheme for 3D holoscopic intra-frame coding through an improved spatial prediction method that explores the particular arrangement of 3D holoscopic images, notably the intrinsic self-similarity of this type of images, without requiring explicit knowledge of how the micro-images are arranged and their actual size.

In the past, the idea of exploiting the non-local spatial correlation has also been tried for 2D images and video in order to further enhance the performance of H.264/AVC intra prediction. Examples are the intra displacement compensation technique introduced in [14], and the method called template matching prediction proposed by T. K. Tan *et al.* [15].

The remainder of the paper is organized as follows. Section 2 briefly explains the general concepts and structure of 3D holoscopic images, in order for the reader to better understand how the spatial prediction will be made. Section 3 describes the proposed improved spatial prediction scheme, while Section 4 performs the evaluation of this scheme. Finally, Section 5 concludes the paper.

2. SPATIAL RELATIONSHIPS IN 3D HOLOSCOPIC IMAGES AND VIDEO FRAMES

G. M. Lippmann first proposed 3D holoscopic imaging in 1908 [4]. With this technology, users can enjoy the impression of depth with full motion parallax without having to wear any special glasses. This is achieved with an array of small spherical micro-lenses, known as a "fly's eye" lens array, which is used to both record and display the 3D holoscopic image (see Figure 1). This array is designed so that, at the display side, different images will be visible depending on the viewing angle. This occurs because only one pixel from each micro-image is visible by the user, which depends on the user's position. With this technology, a simple flat panel display can be used for creating 3D images. If, instead of still images, motion pictures are considered, this leads to 3D holoscopic video. At the capture side, the micro-lens array is applied directly to the sensor, but the basic operation is the same.

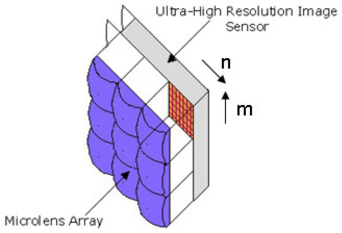


Figure 1 - Micro-lens array applied to an imaging sensor for capturing 3D holoscopic content.

The planar intensity distribution projected behind the micro-lens array, which represents a 3D holoscopic image, consists of a simple 2D array of micro-images of m×n pixels. This happens because of the structure of the micro-lens array that is used for capturing such 3D holoscopic content. As such, this 2D array could be simply encoded by any 2D image or video encoder. However, each microlens can be viewed as an individual small low-resolution camera, recording a different image of the same scene, at slightly different angles. Due to the small angular disparity between adjacent microlenses, a significant cross-correlation exists between neighbouring micro-images. Therefore, this inherent cross-correlation of 3D holoscopic images can be exploited for improving coding efficiency. Additionally, a significant correlation also exists between neighbouring pixels within each micro-image. These two types of correlation are clearly illustrated in Figure 2.

Unidirectional holoscopic imaging is a special case of what has just been described, where a 1D cylindrical lens array is used for both capture and display. In this case, the resulting 3D holoscopic images only exhibit horizontal parallax, with the captured images consisting of one pixel wide vertical lines, where each vertical line behind each cylindrical lens corresponds to one viewing position.

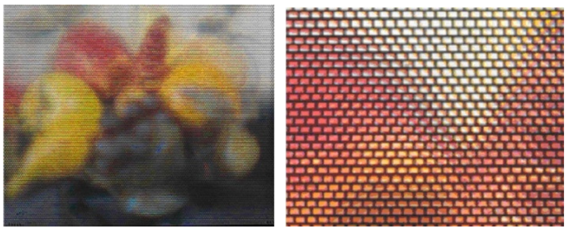


Figure 2 - Example of a 3D holoscopic image: (left) *Fruit* test image; (right) test image enlargement, showing the inherent holoscopic spatial structure due to the micro-lens array.

Although the proposed spatial prediction approach described in the next section is especially suited for full-parallax holoscopic content, it can also be applied when only unidirectional parallax is considered.

3. IMPROVED SPATIAL PREDICTION SCHEME BASED ON SELF-SIMILARITY COMPENSATION

The type of spatial correlation present in 3D holoscopic content can be seen as a kind of self-similarity. As such, it is something that an encoder should exploit in order to improve the coding efficiency.

The use of the term self-similarity in this paper should not be confused with the self-similarity in the context of fractals (and fractal-based image coding), where it basically refers to scale-invariance [16]. In the context of this paper, self-similarity refers only to translations.

3.1. Proposed Codec Architecture

The proposed spatial prediction scheme corresponds to a module that fits in the architecture of a 3D holoscopic image/video codec, as depicted in Figure 3, which is based on the H.264/AVC architecture [17]. This scheme introduces a new set of spatial prediction modes, in addition to all the existing H.264/AVC Intra modes. The new Intra modes (INTRA_SS 16×16, INTRA_SS 16×8, INTRA_SS 8×16, INTRA_SS 8×8 and INTRA_SS SKIPPED) use a self-similarity estimation and a self-similarity compensation block, to find the best predictor for the current block partition being encoded in the area of the current image/frame that has already been encoded (and reconstructed, since this is what will be available at the decoder). In order to avoid possible confusions with traditional spatial prediction done in H.264/AVC codecs, the type of prediction being proposed here will be referred to in the remainder of the paper as self-similarity prediction.

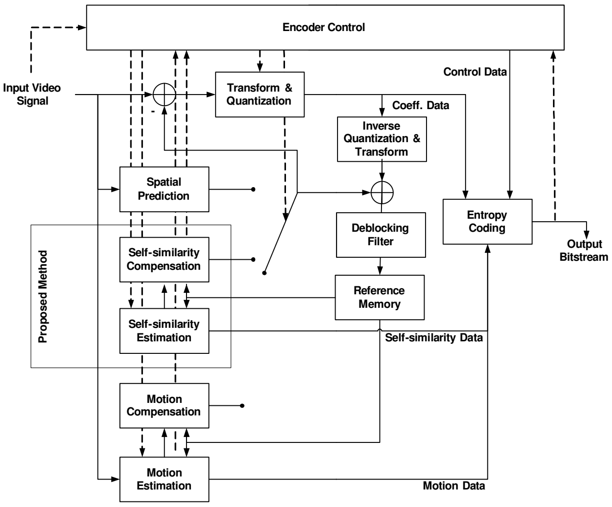


Figure 3 - Proposed 3D holoscopic codec architecture, including the new self-similarity spatial prediction scheme.

3.2. New Self-Similarity Spatial Prediction (INTRA_SS) Modes

Each new mode INTRA_SS 16×16, INTRA_SS 16×8, INTRA_SS 8×16, and INTRA_SS 8×8 specifies a different way to partition the macroblock. For a given macroblock partition with a size of 16×16, 16×8, 8×16 or 8×8 to be encoded, the self-similarity estimation module uses block-based matching (with the same size), in an area previously coded and reconstructed of the current picture, to find a full-pel region with the best match, in terms of the Sum of Absolute Difference (SAD), for prediction of the current macroblock partition. For the INTRA_SS 8×8 mode, each 8×8 macroblock partition can be further split into 8×4, 4×8 or 4×4 sub-partitions for self-similarity estimation, which is indicated in the bitstream with a sub-macroblock type syntax element. The allowed search area for self-similarity estimation is illustrated in Figure 4.

The chosen 16×16, 16×8, 8×16 or 8×8 region that corresponds to the best match becomes the candidate predictor for the current macroblock partition being encoded. The relative position between the current partition (or sub-partition) being encoded and its predictor is here referred to as a self-similarity vector (similarly to a motion vector).

At the start of self-similarity estimation, the INTRA_SS SKIPPED prediction mode, considering a 16×16 block size, is also evaluated as a candidate prediction mode, in which a self-similarity vector is derived from the self-similarity vectors of three previously coded macroblock partitions (left, above and above right), according to the rules defined in H.264/AVC for the INTER_SKIPPED mode [17].

When either the INTRA_SS 16×16, INTRA_SS 16×8, INTRA_SS 8×16 or INTRA_SS 8×8 prediction mode becomes the best mode, in a rate-distortion sense, from the set of all possible intra-coding modes, a distinct self-similarity vector is encoded and transmitted for each macroblock partition or sub-partition. Up to 16 self-similarity vectors can be encoded (i.e., for the INTRA_SS 8×8 mode, when each 8×8 macroblock partition is further partitioned into 4×4 sub-partitions). In addition to the self-similarity vectors, the prediction residual is also encoded and transmitted.

On the other hand, when the INTRA_SS SKIPPED is selected as the best mode (also in a rate-distortion sense), only a skipped macroblock indicator is transmitted.

In the self-similarity compensation block, the inverse quantized and inverse transformed prediction residual is added to the predictor to form the reconstructed macroblock partitions that are stored in the prediction memory, in order to be used for the prediction of future macroblock partitions.

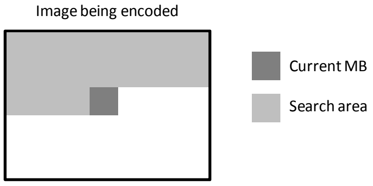


Figure 4 - Illustration of the allowed search area for the estimation of the self-similarity vector.

3.3. Mode Decision

The decision to choose the best macroblock mode is accomplished through the rate-distortion optimization (RDO) technique, where the best macroblock mode is selected by minimizing the following Lagrangian cost function:

$$J_{MODE} = D(MODE, QP) + \lambda_{MODE} \times R(MODE, QP) \qquad (1)$$

where MODE is one of the allowed Intra macroblock coding modes (i.e., INTRA 16×16, INTRA 8×8, INTRA 4×4 or the new INTRA_SS 16×16, INTRA_SS 16×8, INTRA_SS 8×16, INTRA_SS 8×8 and INTRA_SS SKIPPED,), QP is the macroblock quantization parameter, and $D(MODE, QP)$ and $R(MODE, QP)$ are, respectively, the distortion (between the original and the reconstructed macroblock) and the number of encoded bits that will be generated by applying the corresponding MODE and QP. λ_{MODE} is the Lagrange multiplier parameter and is computed as in [13].

Therefore, to encode a given macroblock, the proposed scheme computes the best Intra mode RD cost, J_{INTRA} , from the set of all possible Intra modes by using Equation (1), where

$$J_{INTRA} = \min_{MODE \in \mathbf{S}_{INTRA}} \left(J_{MODE} \right) \qquad (2)$$

where $\mathbf{S}_{INTRA}$ is the set of allowed Intra modes. The best Intra mode is the one with the lowest Intra RD cost.

3.4. Mode Encoding

The proposed codec uses similar syntax elements to those of H264/AVC to encode (and decode) a single macroblock. As can be seen from the macroblock layer syntax depicted in Figure 5, a syntax element called *mb_type* is used to choose between the INTRA (16×16, 8×8 or 4×4) and the new INTRA_SS (16×16, 16×8, 8×16 or 8×8) modes. In the case of the INTRA_SS 8×8, an additional syntax element *sub_mb_type* indicates which further sub-macroblock partition will be used (8×8, 8×4, 4×8 or 4×4). The INTRA_SS SKIPPED mode is signalled by the *mb_skip_flag*, which can be used in alternative to the *mb_type* syntax element. Since the self-similarity vectors are differentially encoded, the *ssvd* syntax element specifies the difference between a given self-similarity vector component to be used and its prediction, which is defined for each macroblock partition or sub-macroblock partition. The method of forming the prediction for different partition sizes is the same defined in H.264/AVC [17]. The *coded_block_pattern* is used to signal which transform blocks have non-zero coefficients, *mb_qp_delta* syntax element indicates a change in the quantization parameter (QP) and, finally, *residual* represents the syntax elements used to encode the residual data.

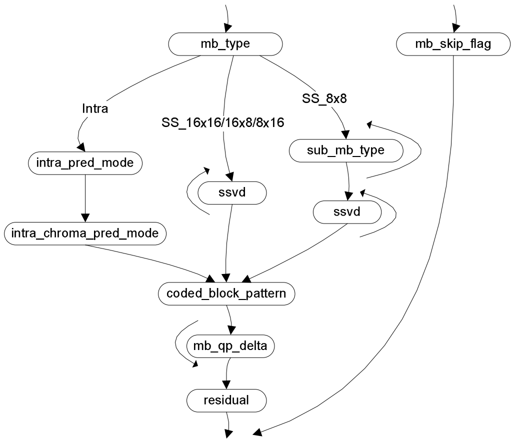


Figure 5 – Macroblock layer syntax elements for proposed codec.

These syntax elements are entropy coded using Context-based Adaptive Binary Arithmetic Coding (CABAC). In this arithmetic coding method, different probability models are maintained for each syntax element (i.e., the context) and each of these models is identified by a context index. The various probability models (or context models) are then updated along time based on the statistics of previously coded data. A detailed description of CABAC can be found in [18].

Before encoding the first macroblock of a slice, it is necessary to perform an initialization process for the context indices. In this process, different contexts are initialized with a pre-computed initial distribution, covering all syntax elements described for the current slice type.

A modified I-slice is defined in order to associate the new INTRA_SS modes and their syntax elements (see Figure 5) with a probability model, consequently the context indices are initialized with the values defined in H.264/AVC for context variables of P-slices [19], due to their similarities with Inter P SKIP, 16x16, 16x8 and 8x8 modes.

The CABAC process for encoding a single syntax element involves three fundamental steps:

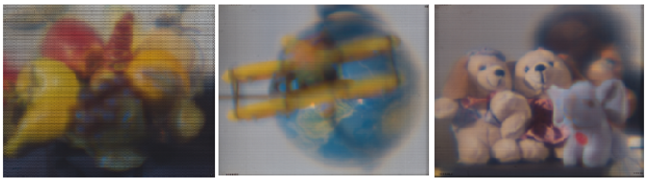
- In the first step, so-called binarization, the syntax element is converted to a series of binary decisions, referred to in the following as *bin*. For each new *mb_type* and *sub_mb_type*, a *bin* is also defined with the specification given in H.264/AVC for P-slices [19], making the correspondence between partition sizes of INTRA_SS and Inter P modes.
- In the context modelling step, for each *bin*, a context index is derived that corresponds to the selection of a context model. Then, the related context model is updated to increment the context index of the *bin* that was encoded.
- In a final step, the *bin* and its associated context model is passed to a binary arithmetic encoding engine.

These proposed adjustments to the initialization and binarization processes make it possible to take advantage of the superior performance of CABAC [18] with respect to other entropy coding algorithms and enables future improvements to be incorporated on the codec proposed here by developing new context models for 3D holoscopic content.

4. PERFORMANCE EVALUATION

This section evaluates the performance of a modified H.264/AVC codec with the additional proposed INTRA_SS modes against the normative H.264/AVC codec using typical settings for intra-coding. In these tests, the deblocking filter has been disabled.

For these tests, the *Fruit*, *Plane* and *Toys* holoscopic test images, with full parallax and 8×7 micro-image resolution, have been used (see Figure 6). These images were recorded on photographic film, placed behind the microlens array, and then scanned using a high resolution scanner.



(a) (b) (c)

Figure 6 - Test images: (a) *Fruit*; (b) *Plane*; (c) *Toys*.

For the INTRA_SS modes, the 16×16, 16×8, 8×16, 8×8 and SKIPPED modes are currently considered. In terms of search pattern, an exhaustive block-based estimation with full-pel positions was used. This way, in the modified codec, a total of eight Intra coding modes were available to encode each macroblock: INTRA 16×16, INTRA 8×8, INTRA 4×4 and the novel INTRA _SS modes: INTRA_SS 16×16, INTRA_SS 16×8, INTRA_SS 8×16, INTRA_SS 8×8 and INTRA_SS SKIPPED.

Table 1 – Relative Intra mode selection statistics

	Test Image (Resolution)		
MB Coding Mode	*Fruit* (680×562)	*Plane* (698×570)	*Toys* (698×570)
INTRA 16×16	2.20 %	18.18 %	20.71 %
INTRA 8×8	4.46 %	13.13 %	19.07 %
INTRA 4×4	0.32 %	2.90 %	4.92 %
INTRA_SS 16×16	19.38 %	19.32 %	13.51 %
INTRA_SS 16×8	11.95 %	14.20 %	13.38 %
INTRA_SS 8×16	34.75 %	10.54 %	8.46 %
INTRA_SS 8×8	17.25 %	3.28 %	3.54 %
INTRA_SS SKIPPED	9.69 %	18.43 %	16.41 %

Table 1 shows the percentage of each Intra coding mode for each test image encoded with a QP of 32. The macroblock mode decisions were done as described in Section 3.3. As can be seen in Table 1, for the *Fruit* image, the newly proposed INTRA_SS modes are clearly the most frequently chosen (i.e., 93%). For the other two images, the INTRA_SS modes decision percentages are also very significant (i.e., 66% for the *Plane* image and 55% for the *Toys* image). The image quality improvement, measured by the luminance Peak Signal-to-Noise Ratio (PSNR Y), is clearly reflected in the rate-distortion curves presented in Figures 7 to 9, where QP varies from 8 to 48 with increments of 4.

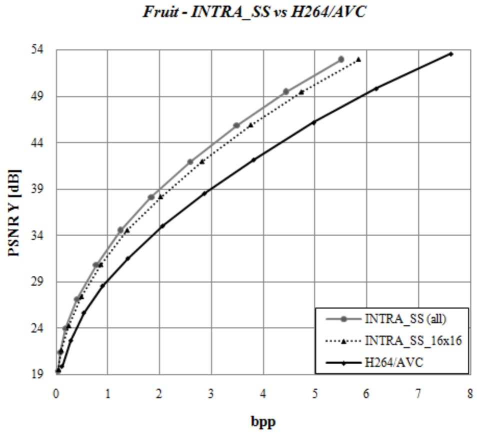


Figure 7 - PSNR results for the *Fruit* image.

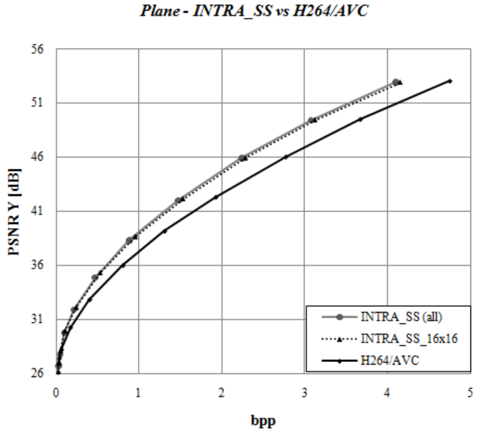


Figure 8 - PSNR results for the *Plane* image.

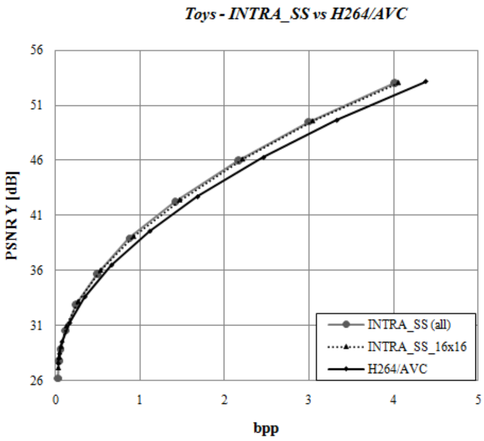


Figure 9 - PSNR results for the *Toys* image.

As can be seen in Figures 7 to 9, the modified codec including the new INTRA_SS modes always outperforms the standard H.264/AVC codec in terms of the achieved PSNR Y values. Moreover, it can also be seen that when all the new INTRA_SS modes are used, better results are achieved than when only the new INTRA_SS 16×16 mode is used. This clearly shows that, in this context, the use of macroblock partitioning is also beneficial. The achieved gains can go up to 5.7 dB for the *Fruit* image, 1.4 dB for the *Plane* image and 1.1 dB for the *Toys* image. These gains are closely related to the relative percentage of INTRA_SS modes usage.

5. FINAL REMARKS

This paper proposes a novel spatial prediction scheme that exploits the self-similarity inherent to 3D holoscopic content. The proposed self-similarity method can be used for still image coding and intra-coding of video. With this scheme, a better spatial prediction was obtained for full-parallax holoscopic content, leading to an improved coding efficiency with respect to H.264/AVC of up to 5.7 dB.

6. ACKNOWLEDGEMENTS

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

7. 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, July 2009.

[2] G. Milnthorpe, M. McCormick, N. Davies, “Computer Modeling of Lens Arrays for Integral Image Rendering”, *Proc. of Eurographics UK Conference*, Leicester, UK, June 2002.

[3] R. Zaharia, A. Aggoun, M. McCormick, “Adaptive 3D-DCT Compression Algorithm for Continuous Parallax 3D Integral Imaging”, *Signal Processing: Image Communication*, vol. 17, no. 3, pp. 231-242, March 2002.

[4] G. Lippmann, “Epreuves Reversibles Donnant la Sensation du Relief”, *Journal de Physique Théorique et Appliquée*, vol. 7, no. 1, pp. 821-825, November 1908.

[5] M. Okui, F. Okano, “3D Display Research at NHK”, *Workshop on 3D Media, Applications and Devices*, Berlin, Germany, October 2009.

[6] A. Aggoun, “A 3D DCT Compression Algorithm For Omnidirectional Integral Images”, *Proc. of the IEEE International Conference on Accoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 2, pp. 517-520, Toulouse, France, May 2006.

[7] R. Zaharia, A. Aggoun, M. McCormick, “Compression of Full Parallax Colour Integral 3D TV Image Data Based on Subsampling of Chrominance Components”, *Proc. of the IEEE Data Compression Conference (DCC 2001)*, pp. 27-29, Snowbird, UT, USA, March 2001.

[8] M. C. Forman, A. Aggoun, “Quantisation Strategies for 3D-DCT based Compression of Full Parallax 3D Images”, *Proc. of the IEE International Conference on Image Processing Applications (IPA 1997)*, Dublin, Ireland, pp. 32-35, July 1997.

[9] A. Aggoun, M. Mazri, “Wavelet-based Compression Algorithm for Still Omnidirectional 3D Integral Images”, Signal, Image and Video Processing, vol. 2, no. 2, pp. 141-153, June 2008.

[10] E. Elharar, A. Stern, O. Hadar, B. Javidi, “A Hybrid Compression Method for Integral Images Using Discrete Wavelet Transform and Discrete Cosine Transform”, *Journal of Display Technology*, vol. 3, no. 3, pp. 321-325, September 2007.

[11] S. Adedoyin, W. A. C. Fernando, A. Aggoun, “A Joint Motion and Disparity Motion Estimation Technique for 3D Integral Video Compression Using Evolutionary Strategy”, *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 732-739, May 2007.

[12] S. Adedoyin, W. A. C. Fernando, A. Aggoun, “Motion and Disparity Estimation with Self Adapted Evolutionary Strategy in 3D Video Coding”, *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, pp. 1768-1775, November 2007.

[13] Joint Video Team, “Joint Multiview Video Coding 8.3.1”, November 2010.

[14] S.L. Yu, C. Chrysafis, “New intra prediction using intra-macroblock motion compensation,” Tech. Rep., JVT-C151, May 2002.

[15] T. K. Tan, C. S. Boon, and Y. Suzuki, “Intra prediction by template matching,” *Proc. IEEE International Conference on Image Processing*, Atlanta, USA, pp. 1693–1696, Oct. 2006.

[16] B. Mandelbrot, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension", Science, New Series, vol. 156, no. 3775, pp. 636-638, May 1967.

[17] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, A. Luthra, “Overview of the H.264/AVC Video Coding Standard”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, July 2003.

[18] D. Marpe, H. Schwarz, T. Wiegand, “Context-based adaptive binary Arithmetic Coding in the H.264/AVC video compression standard”. *IEEE Transactions on Circuits and Systems for Video Technology*, , vol. 13, pp. 620-636, no. 7, July 2003.

[19] ITU-T Recommendation H.264 (2010): Advanced Video Coding for Generic Audiovisual Services, March 2010.