Parameterized Estimation of Common Motion for Image and Depth Sequences

Modem Sudhakar
Dept. Of Electrical Engineering
IIT Hyderabad
Andhra Pradesh, India-502205
Email: ee10m04@iith.ac.in

Kiran Kumar Vupparaboina
Dept. Of Electrical Engineering
IIT Hyderabad
Andhra Pradesh, India-502205
Email: ee11p012@iith.ac.in

Soumya Jana
Dept. Of Electrical Engineering
IIT Hyderabad
Andhra Pradesh, India-502205
Email: jana@iith.ac.in

ABSTRACT

***Abstract*—Representation of 3D video using reference image and depth map sequences (2D+depth) has become standard. Although standards recommend separate encoding of image and depth, exploitation of their correlation opens up opportunities, especially for low-bit-rate applications. To this end, a common motion vector is sometimes used to compensate both the sequences. In this context, we propose a parametric method for estimation of common motion that improves upon the current image-based approach. Specifically, we identify statistical redundancy between image and depth sequences, demonstrate how the said redundancy cannot fully be exploited by a technique based on image sequences alone, and then propose a method to remove such redundancy in an optimized manner. Finally, we demonstrate the efficacy of our technique using well-known stereoscopic video sequences.**

KEY WORDS

3D Video Coding, Depth Image based Rendering, Common Motion Vector, MPEG-2.

I. Introduction

With the growing ubiquity of stereoscopic video, its compression has been receiving considerable attention. Stereoscopic video consists of synchronized sequences of left and right images, which generally exhibit high correlation. In response, state-of-the-art encoders create reference images and corresponding pixel-wise depth information (2D+depth), which exhibit less correlation [1]. Accordingly, video coding standards, including MPEG [2], [3], recommend that image and depth sequences be encoded and decoded separately as shown in Fig. 1. Although simple, such separate encoding prevents the exploitation of image-depth correlation, which, albeit less compared to the left-right image correlation, could still be significant. In this paper, we propose a parameterized motion estimation method that exploits image-depth correlation, while maintaining the essential simplicity of the standard encoder, and achieves significant performance gains at low bit rates. Consequently, our technique assumes significance for limited-bandwidth applications such as mobile 3D-television.

Exploitation of depth-image correlation has already been attempted. One such attempt involves motion estimation in three dimensions [4], which not only is incompatible with current standards based on 2D motion, but also adds to the bandwidth to encode a third motion component. In contrast, common motion for image and depth map sequences has been proposed by Grewatsch *et al.*; however, such common motion is estimated based on the image sequence alone [5]. Consequently, certain residual redundancy remains unless image motion and depth motion are perfectly correlated. Later, De Silva *et al.* added the flexibility of estimating a depth (z-) component of motion vector from depth sequences [6]. Benefit of their approach remains unclear, as such z-component is essentially the dc value of the depth error block, and handled by the usual transform coder using discrete cosine transform (DCT). In our paper, we also assign a common motion vector to both image and depth blocks, but do not estimate it based on image sequences alone. Instead, we combine image error and depth error according to a parametric weight, and pick the common motion that minimizes the combined error. Subsequently, we optimize the parameter to achieve a certain rate-distortion goal. Interestingly, our method coincides with the image-based approach for perfectly correlated image and depth motion. However, for practical cases with imperfect correlation, our method exhibits superior performance. We provide an intuitive explanation below, and empirical corroboration later in the paper.

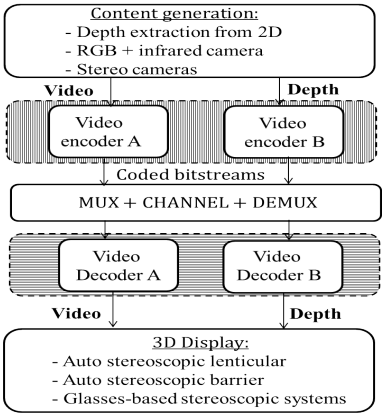
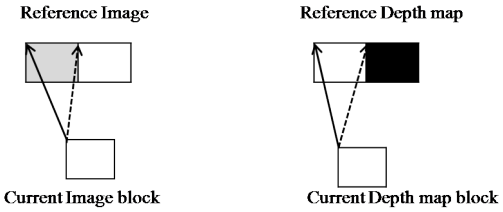


Fig. 1: Separate encoding of video and depth sequences [2].



(a) (b)

Fig. 2: Parameterized motion: (a) image, and (b) depth.

Let us turn to an extreme scenario of imperfect correlation depicted in Fig. 2a. Specifically, consider two contiguous blocks in the reference frame (upper row): the left (resp. right) block in image (a) corresponds to the left (resp. right) block in depth (b). Note that image pixels in left block are nearly white (value $= 250$, say), and those of right block are white (value $= 255$), while depth values for left block are white (255, i.e., very near), and that for the right block is black (0, i.e., very far). Now, assume a hypothetical situation where the right block approaches the camera very fast while getting shrunk so that its image pixel values remain unchanged, and the field of view it occupies does not change. Specifically, refer to the lower row of Fig. 2a, where the current image block remains white (255), but the current depth map goes from black to white (255). Now, if one performs motion estimation separately for image and depth, clearly the matched image block will be the right one, whereas the matched depth block would be the left one.

In other words, the image motion vector would completely misrepresent the depth motion, and use of that vector as common motion would produce high depth error alongside zero image error. On the other hand, if one uses the depth motion vector as common motion, in this example one would obtain low image error alongside zero depth error. Interestingly, our parametric approach would pick the depth-based motion in this case, and perform better than the fixed image-based method. More generally, via a parameter our technique has the flexibility of choosing either of the two extremes of image-based and the depth-based methods, or an intermediate scenario where image error and depth error both contribute. In fact, we shall see that the optimal parameter for real stereoscopic sequences are generally neither of the extremes, leading to appreciable improvement in performance.

Rest of the paper is organized as follows. Section II presents traditional 2D and 3D video coding. Parameterized common motion estimation is proposed in Section III. Section IV discusses experimental results. Finally, Section V concludes with a discussion.

II. Background: 3D video compression

A. 2D Video Compression

3D video (2D+depth) compression standards extend 2D compression as shown in Fig. 1, and elaborated below [3]. A 2D video sequences is divided into groups of pictures (GOP), each group consisting of one Intra-coded (I) frame, several Predictively-coded (P) frames and a larger number of Bi-directionally predictively-coded (B) frames. A typically GOP looks like "IBBPBBPBBP". The temporal redundancy is removed using motion compensation in non-intra (i.e., P and B) frames, whereas the spatial redundancy is removed in all frames by means of transform coding. Finally, the motion vectors, quantized prediction errors, and other overhead data are entropy encoded using variable-length codes (VLC).

Specifically, each frame is divided into 16×16 macroblocks (MB), and motion of the current MB is estimated using block-matching algorithm. For example, using the "sum of absolute differences" (SAD) criterion, the estimated motion vector is given by

$$\underline{v}^* = \arg\min_{\underline{v}} \|EB(\underline{v})\|_1, \tag{1}$$

where the error macroblock is denoted by $EB(\underline{v}) = MB - \widehat{MB}(\underline{v})$, MB indicates current macroblock, $\widehat{MB}(\underline{v})$ indicates the macroblock in reference frame with motion vector $\underline{v}$, and $\|\cdot\|_1$ denotes the l_1 norm. The error macroblock $EB(\underline{v}^*)$ is transform-coded using dicrete cosine transform (DCT) and quantized. Different data components such as motion vectors and quantized error are entropy-coded using huffman, run-length and other coding schemes.

B. 2D+Depth Compression

State-of-the-art 3D video compression makes use of Depth-Image-based-Rendering (DIBR), where virtual left and right views could be synthesized from a reference image and its perspective depth map (2D+depth), and *vice versa* [1]. The mathematical relation between reference image point (x, y) and virtual left and right views, (x_L, y) and (x_R, y), respectively, are given by

$$x_{L/R} = x \pm \frac{f_x.d}{2}\left(\frac{1}{z(x,y)} - \frac{1}{l}\right), \tag{2}$$

where take '$+$' for x_L and '$-$' for x_R in '$\pm$', and denote by d the base line length, by $z(x, y)$ the depth value for pixel (x, y), by f_x the focal length of the reference camera, and by l the furthest representable depth. Finally, the absolute depth $z(x, y)$ is nonlinearly mapped to 'depth image' $D(x, y)$,

$$D(x,y) = 255\left(\frac{1}{z(x,y)} - \frac{1}{z_{max}}\right) / \left(\frac{1}{z_{min}} - \frac{1}{z_{max}}\right), \tag{3}$$

which takes values 0-255 (the minimum depth z_{min} maps to '255', and maximum depth z_{max} to '0').

III. Parameterized Estimation of Common Motion

Current standards, such as MPEG, recommend separate encoding of (hence, separate motion vectors for) image and depth sequences as shown in Fig. 1 [3], which does not exploit the correlation between image and depth motion. In contrast, use of image motion vectors for depth is reported to provide improvement in certain rate regimes [5], [6]. In this backdrop, we note that the image motion as representative of depth motion is guaranteed to be efficient only when both types of

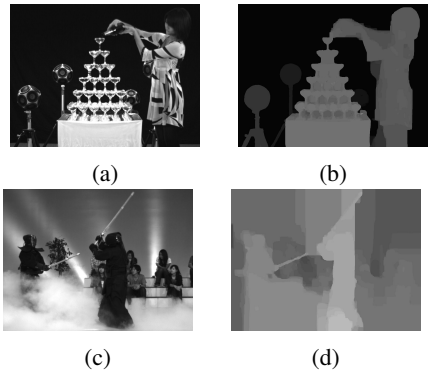


Fig. 3: Champange Tower sequence: image (a) and depth (b); Kendo sequence: image (c) and depth (d) [7].

motion are perfectly correlated, and propose a parameterized technique for selecting representation motion for both image and depth sequences. Our technique reduces to the image-based method for perfectly correlated image and depth motion, and outperforms image-based techniques in general (refer Fig. 2).

A. Motion Estimation: State-of-the-art

1) Separate Motion Estimation for Image and Depth: Image and depth sequences are treated as two separate video sequences [3], and motion estimation is performed in each using traditional block matching algorithm. Consequently, each macroblock location is assigned two motion vectors, one for the image and one for the depth sequence.

2) Motion Estimation using Image Sequence Only: Image motion is estimated using block-matching algorithm as usual. However, this motion vector is used for motion compensation not only for image sequences but also for depth sequences [5], [6]. In other words, each macroblock location is assigned only one motion vector, whereby saving about half the motion bits. Further, the resulting image error data is same as that in case of separate encoding. However, due to motion mismatch, depth error data are generally more voluminous compared to that in separate encoding. However, due to substantial correlation between image and depth motion, the increase in depth error volume has been shown to be dominated by the savings in motion bits for equivalent distortion performance in certain rate regimes leading to higher encoding efficiency.

B. Proposed Method: Parameterized Motion Estimation

Despite its success, the aforementioned image-based technique has unexploited redundancy. To demonstrate such redundancy, an extreme scenario has been constructed in Fig. 2, where a depth-based motion proves to be a more efficient representative motion for both image and video sequences. In view of this, we propose for estimating a representative motion a parameterized convex weight ($\lambda \in [0,1]$) on image and depth errors, encompassing both image-based ($\lambda = 1$) and depth-based ($\lambda = 0$) approaches. As opposed to fixing an ad hoc weight a priori (e.g., $\lambda = 1$ for image-based scheme),

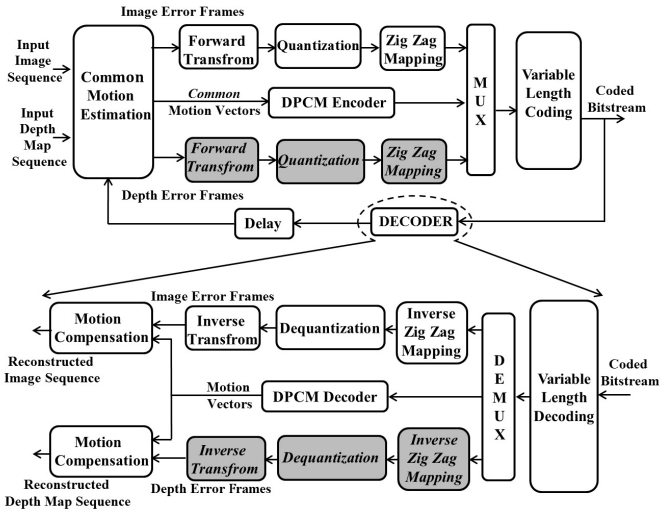


Fig. 4: Schematic of Proposed System

this further allows us the flexibility of choosing an optimized weight λ according to suitable rate-distortion criterion. Note that we need to modify Fig. 1 only slightly to obtain a 2D+depth encoder for our method as depicted in Fig. 4.

1) Parameterization and Motion Estimation: Denote by $EB_I(\underline{v})$ and $EB_D(\underline{v})$, respectively, the co-located image and depth error blocks for common motion vector $\underline{v}$. Then the weighted sum of absolute differences (SAD) is given by $\lambda\|EB_I(\underline{v})\|_1+(1-\lambda)\|EB_D(\underline{v})\|_1$, where parameter $\lambda \in [0,1]$ determines the convex weight, and $\|\cdot\|_1$ indicates the l_1-norm. For a given $\lambda \in [0,1]$, the estimated common motion vector is given by

$$\underline{v}^*(\lambda) = \arg\min_{\underline{v}} \; [\lambda\|EB_I(\underline{v})\|_1 + (1-\lambda)\|EB_D(\underline{v})\|_1], \quad (4)$$

under the SAD criterion. One often normalizes the SAD by the number of pixels in a macroblock (i.e., by $256 = 16\times 16$), and report the mean absolute difference (MAD). In general, estimation motion vector $\underline{v}^*$ varies with λ.

2) Optimization of parameter $\boldsymbol{\lambda}$: Further, we have the flexibility of choosing λ in a rate-distortion-optimal manner as described below. Assuming a fixed quantization matrix, denote the aggregate bit-requirement for encoding error-blocks $EB_I(\underline{v}^*(\lambda))$ and $EB_D(\underline{v}^*(\lambda))$ by $R(\lambda)$, and the average distortion due to left and right views by $D(\lambda)$. Here the above views are generated from motion-compensated image and depth frames with common motion vector $\underline{v}^*(\lambda)$. Further, find the minimum bit-rate

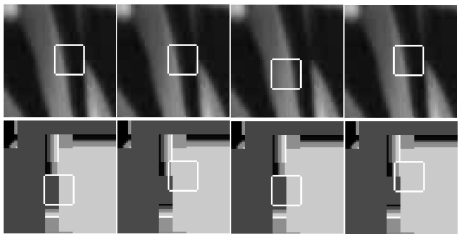
$$R_{min} = \min_{\lambda\in[0,1]} R(\lambda) \quad (5)$$

over λ. Next, for a tolerance level $\delta > 0$, set target rate

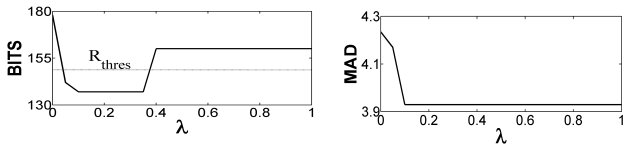
$$R_{thres} = (1+\delta)R_{min}. \quad (6)$$

Finally, obtain optimal λ^* that minimizes the distortion, subject to target rate R_{thres}, i.e.,

$$\lambda^* = \arg \min_{\lambda: R(\lambda)\le R_{thres}} D(\lambda). \quad (7)$$



(a)



(b) (c)

λ	0	.05	.1 to .35	.4 to 1
v_x	0	-8	-8	-8
v_y	-8	-5	-3	-2

(d)

Fig. 5: Selection of parameter λ for macroblock (56,73) in frame number 113 for Champagne Tower sequence [7]: (a) Upper row corresponds to image, and the lower row to depth. The four columns identifies the reference block for motion estimation in four respective modes - separate, image-based, depth-based, and parameterized; (b) $R(\lambda)$ versus $\lambda \in [0, 1]$, (c) $D(\lambda)$ versus $\lambda \in [0, 1]$ and (d) tabulation of motion vectors $(\underline{v}^*(\lambda))$ for various values of $\lambda \in [0, 1]$.

Finally, in view of (4), the estimated motion vector is declared as $\hat{\underline{v}} = \underline{v}^*(\lambda^*)$.

3) Illustration of Parameter Optimization: Optimal selection of parameter λ is illustrated in Fig. 5 in relation to traditional techniques. Specifically, we take the Champagne Tower sequence [7], and consider the macroblock indexed by (56,73) (i.e., its upper left pixel has the coordinate $((56-1)\times 16+1, (73-1)\times 16+1)$ in frame number 113. In Fig. 5(a), the upper row corresponds to image, and the lower row to depth. Further, the first column corresponds to separate motion estimation for image and depth; the second column corresponds to motion estimation based on image alone, and use of estimated image motion as depth motion; third column is analogous to the second column with the role of image and depth reversed; and the fourth column corresponds to the proposed parameterized motion estimation. Further, for different values of $\lambda \in [0, 1]$, the variation of $R(\lambda)$ (while identifying R_{thres}) and $D(\lambda)$ are plotted in Figs. 5(b) and 5(c), respectively. Finally, the estimation motion vectors $\underline{v}^*(\lambda)$ (with components v_x and v_y) are tabulated for various values of λ in Fig. 5(d). Note that all values of λ in the range 0.1–0.35 are optimal because those lead to both the least bit-rate as well as the least distortion. As a representative, we pick $\lambda^* = 0.1$, and the estimated motion vector turns out to be $\hat{\underline{v}} = \underline{v}^*(\lambda^*) = (-8, -3)$. Note that the image-based motion estimation (corresponding to $\lambda = 1$) would produce an estimated motion vector equal to $(-8, -2)$, instead.

C. Perceptually Motivated Distortion Metrics

In practice, 3D video content is meant for human perception where a distortion criterion such as MAD may not be appropriate . Noting this, various perceptually motivated objective measures have been proposed. Examples of such measures with varying sophistication include Structural Similarity index (SSIM) [8], and Depth Quality metric (DQM) [9]. Of the above, the SSIM in the 3D context is given by

$$SSIM = \frac{S(x_l, y_l) + S(x_r, y_r)}{2} \tag{8}$$

i.e., the average of SSIM's for left and right views. Here the SSIM of either view is defined by

$$S(x, y) = \frac{4\mu_x\mu_y\sigma_{xy}}{(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_y^2)} \tag{9}$$

based on quantities $-$ x-mean μ_x, y-mean μ_y, x-variance σ_x^2, y-variance σ_y^2, and cross-correlation σ_{xy} [8]. On the other hand, the DQM, defined by

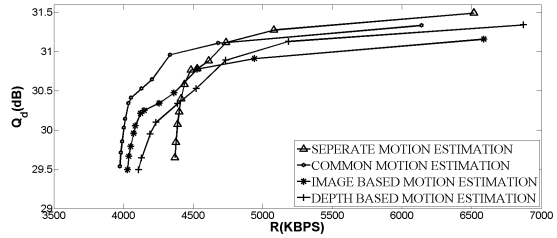
$$Q_d = 10\log_{10}\left(\frac{255^2 N_1 \times N_2}{\sum_{(x,y)} E_R(x, y)}\right) \tag{10}$$

is specifically proposed for 3D video, where $N_1 \times N_2$ indicates the depth map size, $E_R(x, y) = E_1(x, y) + E_2(x, y)$ denotes the view rendering error for pixel (x, y). Specifically, a 3D position P, to be ideally rendered to (x, y), could be misrepresented on both left and right views due to quantization leading to left and right rendering errors $E_1(x, y)$ and $E_2(x, y)$. In our paper, we report our results in terms of SSIM and DQM.

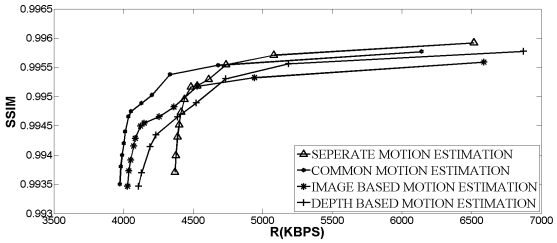
IV. Results

In this section, we demonstrate the advantage of parameterized motion estimation using two stereoscopic sequences: Champagne Tower, and Kendo (see figure 3) [7]. Specifically, for each sequence, one GOP of ten frames of the form IBBPBBPBBP is considered. Further, four schemes are considered, namely, (i) separate motion estimation for image and depth, (ii) common motion estimation based on image alone, (iii) common motion estimation based on depth alone, and (iv) common motion estimation based on the proposed parameterized strategy. These schemes are then compared from the rate-distortion perspective.

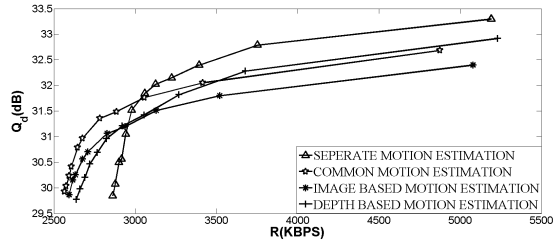
Specifically, in Fig. 6, we observe that the proposed parameterized technique performs consistently better than the other techniques at lower rates, while separate motion estimation outperform the rest at higher rates. This result is not surprising, because at low bit-rates one would expect error to be encoded with very few bits, but bits required for motion data would not alter very much. Thus a common-motion technique, which saves about half the motion data, would be efficient because



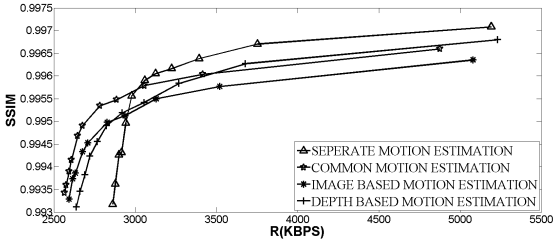
(a) Rate-Q_d plot for Champagne Tower sequence.



(b) Rate-SSIM plot for Champagne Tower sequence.



(c) Rate-Q_d plot for Kendo sequence.



(d) Rate-SSIM plot for Kendo sequence.

Fig. 6: Rate-Distortion (SSIM/Q_d) plots for Champagne Tower and Kendo sequences [7].

the already high error cannot worsen appreciably. On the other hand, at high bit rates separate motion estimation could be more targeted than any common motion. Finally, at low rates, our parameterized approach proves to be more efficient that image-based (as well as depth-based) common motion. For instance, Q_d-improvements of 0.55dB at 4072Kbps, and 0.45 db at 2651 kbps, over image-based technique are observed for Champagne Tower and Kendo sequences, respectively (using linear interpolation Fig. 6).

Table I provides a closer look at the number of bits required to encode motion and error data in various approaches. Clearly, common motion (both in image-based and parameterized cases) achieves significant savings in motion bits, while conceding relatively small increase in number of bits in error data. Further, the proposed parameterized approach allocates about

Champagne Tower	Error$_P$ (Image+depth)/ common	Error$_B$ (Image+depth)/ common	M.V.$_P$ (Image+depth)/ common	M.V.$_B$ (Image+depth)/ common
Seperate	306431+288436	577675+576004	149233+81723	301127+163516
Image Based	315691+294847	579360+580823	149020	300289
Proposed	312838+288732	580412+576048	147507	271504
Kendo Sequence				
Seperate	217253+196628	376025+372294	100184+69237	209845+154420
Image Based	238546+211069	383369+389276	100503	208942
Proposed	230426+203263	382579+376337	103013	200914

TABLE I: Number of bits allocated for error and motion data for P and B frames in a GOP under constant distortion for Champagne Tower (Q_d=29.7 db) and Kendo (Q_d=30.5 db) sequences.

the same number of bits to motion data as in the purely image-based approach. However, the former saves on the error bits overall. Notice that our method increases the volume of image error data in certain cases, however, such increase is more than compensated by the reduction in the volume of depth error data.

V. Discussion

In this paper, we proposed a novel parameterized estimation method for estimating common motion, and demonstrated its advantages in the low-rate regime. We believe that bandwidth limited (especially, mobile) applications could benefit from our technique. Another attractive feature arises from its conformity with standard codecs. In future, we plan to test our ideas on reference implementations of various video coding standards such as MPEG-2, H.264/AVC and H.265 to establish wider applicability.

References

[1] C. Fehn,"Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," Proc. of SPIE Stereoscopic Displays and Virtual Reality Systems XI,pp. 93–104, San Jose, CA, USA, Jan. 2004.

[2] A. Bourge, J. Gobert, F. Bruls, "MPEG-C part 3: Enabling the introduction of video plus depth contents," in Proc. of IEEE Workshop on Content Generation and Coding for 3D-television, Eindhoven, Netherlands, Jun. 2006.

[3] ISO/IEC JTC 1/SC 29/WG 11. Committee Draft of ISO/IEC 23002-3 Auxiliary Video Data Representations. WG 11 Doc. N8038. Montreux, Switzerland, Apr. 2006.

[4] B. Kamolrat, W.A.C. Fernando, M. Mrak,"3D motion estimation for depth information compression in 3D-TV applications," Electronic Letters , vol. 44, no. 21, Oct. 2008.

[5] S.Grewatsch, E.Muller"Sharing of Motion Vectors in 3D Video Coding," Proc. of Int. Conf. on Image Processing (ICIP '04), vol. 5, pp. 3271–3274, Singapore, Oct. 2004.

[6] D.V.S.X. De Silva, W.A.C. Fernando, S.L.P. Yasakethu, "Object based coding of the depth maps for 3D video coding," IEEE Trans.on Consumer Electronics, vol. 55, no. 3, pp. 1699–1706, Aug. 2009.

[7] http://www.tanimoto.nuee.nagoyau.ac.jp/ fukushima/mpegftv/

[8] Zhou Wang, Ligang Lu, Alan C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, Volume 19, Issue 2, Feb. 2004, pp. 121–132.

[9] D.V.S.X. De Silva, W.A.C. Fernando, S.T. Worrall, A.M. Kondoz, "A novel depth map quality metric and its usage in depth map coding," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4, May 2011.