Object-based 2D-to-3D Video Conversion for Effective Stereoscopic Content Generation in 3D-TV Applications

Yue Feng[1], Jinchang Ren[2*] and Jianmin Jiang[1]
[1] Digital Media and Systems Research Institute, University of Bradford, Bradford, U.K.
[2] Centre for excellence in Signal and Image Processing, University of Strathclyde, Glasgow, UK
y.feng2@bradford.ac.uk jinchang.ren@eee.strath.ac.uk j.jiang1@bradford.ac.uk

* indicates corresponding author.

***Abstract*−Three-dimensional television (3D-TV) has gained increasing popularity in the broadcasting domain, as it enables enhanced viewing experiences in comparison to conventional two-dimensional (2D) TV. However, its application has been constrained due to the lack of essential contents, i.e., stereoscopic videos. To alleviate such content shortage, an economical and practical solution is to reuse the huge media resources that are available in monoscopic 2D and convert them to stereoscopic 3D. Although stereoscopic video can be generated from monoscopic sequences using depth measurements extracted from cues like focus blur, motion and size, the quality of the resulting video may be poor as such measurements are usually arbitrarily defined and appear inconsistent with the real scenes. To help solve this problem, a novel method for object-based stereoscopic video generation is proposed which features i) optical-flow based occlusion reasoning in determining depth ordinal, ii) object segmentation using improved region-growing from masks of determined depth layers, and iii) a hybrid depth estimation scheme using content-based matching (inside a small library of true stereo image pairs) and depth-ordinal based regularization. Comprehensive experiments have validated the effectiveness of our proposed 2D-to-3D conversion method in generating stereoscopic videos of consistent depth measurements for 3D-TV applications.**

***Index Terms*−Broadcasting, 3D-TV, 2D-to-3D conversion, computer vision, multimedia systems.**

I. Introduction

Stereoscopic or stereo television, also referred to as 3D-TV, can expand users' experiences beyond traditional 2D-TV broadcasting by offering programs with depth impression of the observed scenes [1]. In fact, 3D has been successfully commercialized as stereo movies, such as those by IMAX [2], for people to watch in the cinema, using special devices. Given that the popularity of 3D programs has dramatically increased, 3D-TV has been identified as a possible breakthrough for conventional TV technologies to satisfy the coming need for watching 3D programs at home.

To develop a practical 3D-TV system, an optimized processing chain is required to cover several key parts as illustrated in Fig. 1, including content generation, coding and transmission, decoding and display, and human 3D perception [1, 3, 5]. As shown in Fig. 1, generation of 3D contents is the first step in a 3D-TV system, for which three possible solutions are presented. In the first solution, stereoscopic videos are generated using stereo or multi- camera systems, such as a dual-camera system in [6-7, 10] and a 64-camera system in [9]. In fact, the majority of material available for 3D-TV broadcast today has been produced using a stereo-camera setting, where the left-eye and the right-eye views with slightly different perspectives are separately recorded to form a stereo pair. To avoid geometrical distortions and depth plane curvature, directors and camera operators need to be highly skilled in stereoscopic geometry and camera calibration; thus, this inevitably creates a huge barrier in producing such films [3].

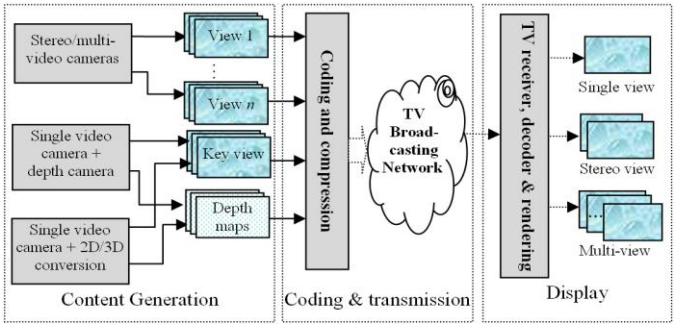


Fig. 1: Diagram of a typical 3D-TV system.

By introducing an infrared range camera, Zcam, as an add-on to existing camera systems, depth information can be captured [1]. In comparison with stereo and multi- camera systems, a depth camera can handle varying conditions more easily [3]. Using depth-image-based rendering (DIBR) techniques, stereoscopic content is generated from one scene image (namely key view) and its associated depth maps [8]. The main challenge here is how to recover occluded regions and how to correct holes and misaligned data in depth maps in synthesizing the left-eye and the right-eye views, such as those reported in [11, 41-43, 51, 54].

In the third solution, the depth camera is replaced by a 2D-to-3D converter, where depth information is extracted from a monoscopic image sequence using computer vision technologies [1]. Firstly, this helps to reduce the overall cost of the system. More importantly, it enables large existing libraries of 2D program material to be reused. As a result, this solution provides a practical way to solve the bottleneck of 3D-TV broadcasting, i.e., lack of program material. Since DIBR is also an essential part in these conversion systems, relevant techniques in recovering occlusions and correcting depth maps are also emphasized for stereoscopic video generation.

Since extraction of depth information from a single sequence is an ill-posed problem in the computer vision domain, 2D-to-3D conversion has two fundamental drawbacks, i.e., inconsistent estimation of depth and recovery of depth ordinal. Inconsistent depth and its ordinal may lead to contradictory signals received by the brain, although human visual perception has been found to be tolerant to certain inaccuracy of estimated depth maps. Since this appearance is contrary to our habitual stereoscopic perception, viewers will feel uncomfortable due to a motion-sickness-like feeling [14].

In this paper, a novel 2D-to-3D conversion method is proposed to solve the two drawbacks above. To determine depth ordinal, occlusion reasoning is applied using cues of bidirectional optical-flow fields. To enable consistent estimates of depth, the disparity value is extracted using object-based matching with a disparity library, followed by regularization-based refinement using determined ordinal. Experimental results have demonstrated the effectiveness and efficiency of the proposed methodology.

Although the proposed method combines our previous work in [38, 45-46], improvements are made for both effectiveness and robustness. Firstly, cues like blurred and un-blurred regions which were used in [45] are not required here to avoid ambiguity in applying blurred regions as out of focus ones because blur can occur from motion. Instead, occlusion reasoning is utilized to help determine depth ordinal. Secondly, in [38] and [46] shape and texture matching are respectively used to estimate the depth map. However, the matching above may produce inconsistent depth values especially when multiple objects exist; here depth-ordinal based regularization is introduced.

The rest of the paper is organized as follows. Section II introduces background information of disparity and depth as well as related work in 2D-to-3D conversion. Section III discusses the process on how to determine depth ordinal using optical-flow based occlusion reasoning. In section IV, technical details regarding object-based determination of depth are presented. Experimental results, evaluations and discussions are given in Section V, followed by some concluding remarks drawn in Section VI.

II. Background and Related work

In this section we describe background information and state of the art approaches in 2D-to-3D conversion. Firstly, the relationship between disparity and depth is presented. Secondly, existing work in extracting depth are categorized into three main classes for discussion, including depth from geometric constraints, depth from focus/defocus analysis and depth from motion. Details are presented below.

A. Depth and Disparity

Human stereoscopic vision relies upon the fact that the viewer acquires the scene from two slightly different projections of the world onto the retinas of two eyes, each from a slightly different viewpoint [4]. The spatial differences in these two images are called disparity. Given the disparity information associated with any pair of images, our brain can generate depth perception by fusing them together. The idea of generating stereo perception from disparity information inspired the capture of stereoscopic video using two cameras with the same setting to take pictures of one scene simultaneously [35, 44, 56]. The two cameras are separated by a distance that is exactly like our eyes. To show the relationship between depth and disparity, the stereo geometry involving the use of a stereo camera is illustrated in Fig. 2.

In Fig. 2, C_l and C_r denote the left- and the right- cameras, o_l and o_r their corresponding image planes. For an object point O, it is captured by the two cameras and forms one pixel on each of two image planes. Let x and x' denote the coordinates of the two pixels, and Z is the depth between object O and the camera plane, i.e., the corresponding depth. Also we define t_c as the distance (baseline) between the left and right cameras, and f as the camera focal length. The disparity, d, which is defined as the distance between x and x', is found inversely proportional to Z as follows [16]:

$$d = |x - x'| = \frac{ft_c}{Z} \tag{1}$$

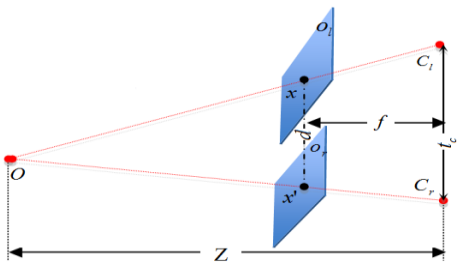


Fig. 2. Stereo geometry for stereo camera

Eq. (1) shows that the disparity of the point O can be decided by knowing the depth and the camera setting values such as the focus length f and the distance t_c. Consequently, in this context, finding the depth is equal to finding the disparity. One possible way to determine the disparity of one point is to compute the distance between the point on the left-image and its corresponding position on the right image. In addition, given one image and its disparity map, a new image can be created by taking the given image as the left-image of a stereo pair and shifting it towards right side to create a right-image according to the determined disparity map.

B. Estimating Depth/Disparity from Monoscopic Sequence

As a well-known problem within the computer vision domain, depth/disparity extraction from a monoscopic image sequence has been investigated for decades. In general, separate or combined cues are used by existing approaches, including shading, texture, blur, motion, geometric perspective, and atmospheric haze [33, 34, 48, 49, 53]. Due to its ill-posed nature, the problem is usually regulated by certain assumption-based constraints. Once the depth map is extracted, it can be widely applied in many broadcasting applications including 3D-TV and postproduction like matting [28]. It is worth noting that conventional methods using shading and texture for 3D scene extraction [13] are seldom used in this context, as they are not guaranteed in generic videos. Typical models and methods for depth extraction in 2D-to-3D conversion are summarized as follows.

1) Depth from geometric constraints

In [12], a DepthFinder system to assist a driver when driving an automobile is reported. The system extracts the depth information of a visual scene from a monocular image sequence using a camera that is mounted on a moving vehicle. The principle behind the depth extraction consists of exploring geometric constraints between camera positions and object images at two moments. With the prior knowledge of camera settings, such as focal length, moving distance of the vehicle is employed as the baseline for depth estimation. In [27] and [31], geometric perspectives like vanishing points and lines are used to establish depth gradient planes. In [50], geometric modeling is applied after foreground detection to form a planar representation of the background, which is similar to that of the depth gradient planes. Finally, intensity-related relative depth is assigned to each gradient plane based on the assumption that a higher depth level corresponds to lower grey values and vice versa.

To enable depth estimation using geometric constraints, controlled environments are required to provide additional information in modeling, such as moving distance in [12] and depth gradient planes in [27, 31, 50]. As a result, this method fails to apply to generic cases.

2) Depth from focus/defocus (blur) analysis

Analysis of focus/defocus, also referred to as blur analysis, is another important method for depth estimation, where the depth in a visual scene is determined by modeling the effect of varying focal parameters on the image [15, 21]. By examining the amount of blur in the image, focal length is obtained as depth through inverse filtering [23], such as by estimating the point spread function (PFS) [45]. In principle, this type of method has two drawbacks. First, blur is not available for the general case. More importantly, blur can be produced by many factors and not just focus length, such as lens aberration, atmospheric interference, and motion [49]. As a result, the application of this method is also constrained.

3) Depth from motion

Based on vision geometry, objects moving horizontally at a similar speed but with a different distance to the camera will produce different results in a recorded sequence, such as the closer the object to the camera, the bigger the change in distance of two continuously frames. Cues from this catalog are occlusion [16], optical flow [17] and motion vector [20], where the occlusion is reflected in the motion and optical flow as a global pattern of retinal motion. Although these cues can be analyzed to produce the depth ordinal for regions which represent the spatial correlation among each object, this approach still cannot obtain sufficient useful depth information to generate a full depth map for the whole frame.

A further study utilizing this principle is reported in [19], in which a modified time difference method (MTD) is applied to detect object motion and estimate the image presentation time delay to create a stereo pair. Similar ideas can also be found in [10, 22]. The results show that MTD works positively in simple cases, but fail for scenes with more complicated motion. In [18], visual depth perception (disparity) is estimated using the horizontal component of motion vectors between consecutive video frames, and further adjusted via a nonlinear model to scale the obtained motion vector. This is implemented using a H.264/MVC based scheme, as accurate motion vectors up to a quarter-pixel can be provided in the standard. Similar work can be found in

[47], where depth is extracted proportionally to the determined motion magnitude, estimated using feature-based and block-based approaches. It is found that the feature-based motion estimation approach outperforms the block-based one, due to the fact that a global optimized estimation can be achieved by prior solution. However, the assumption that a faster moving object will have a larger disparity, i.e., closer to the camera, does not always hold, especially when there are combined camera and object motions. In [37], motion is converted into disparity values with consideration of multi-user conditions and characteristics of the display device. The latter is applied to determine the valid range of depth. After motion estimation, three cues are used to decide the scale factor associated with the motion-to-disparity conversion, i.e., the magnitude of motion, type of camera movements and scene complexity. However, the overall strategy in determining the depth is still heuristic.

Given that motion between two consecutive frames is usually very small and sensitive to illumination changes, these methods are unlikely to produce consistent and satisfactory results [48]. Furthermore, how to deal with occlusions and recover layered motions also need to be solved.

4) Other approaches

In [25] and [26], edge information is used to generate sparse depth maps for DIBR, as the visual system combines the available depth information at the boundary regions with pictorial depth cues to fill in the missing areas. Since the perceived depth is qualitative, this approach is suitable for applications where the accuracy of depth is not crucial [49, 55]. In [32], depth is determined via supervised machine learning, where the input is (r, g, b, x, y) from 2D images and the output is the depth z. Given that depth estimation from monoscopic sequence is an ill-posed problem, learning from a single pixel without support of additional information like shape, motion and shading information cannot generate reasonable results. In [31], color-based image segmentation and heuristic region classification are utilized to assist depth map estimation. With detected vanishing points/lines and classified regions, a combined fusion scheme is introduced to approximate depth map generation. Owing to the limitation of a predefined set of six types of regions, including sky, mountain, land, etc., heuristic rules used by this method seem inadequate in dealing with general cases. In [14], motion detection is used to separate foreground regions from background by examining a pixel's intensity difference in two neighboring frames. The assumption here is that the foreground is composed of near-located moving pixels, and the background contains far-located static pixels. As a result, the intensity difference of the pixel in two neighboring frames is taken as stereoscopic depth. This will inevitably lead to errors, especially when the background contains salient regions like edges and textures or when camera motion exists. In [52], depth is recovered using line tracing followed by an edge-preserving recursive refinement filter. These traced lines will form segmented regions, where each line is assigned with a constant depth determined by its relative height. Three constraints in terms of edge tracing, smoothness and elasticity are utilized to ensure the accuracy of the extracted depth maps. In general, the performance of this method relies on the quality of line tracing, thus it may fail to deal with complex scenes.

C. Summary

Due to its high flexibility and low cost, 3D video generation from 2D video has become increasingly important for DIBR-enabled 3D-TV applications [11]. Among the methods discussed above, depth from motion appears to be one step ahead of other techniques as it can solve real life problems. On the contrary, others can only be applied in limited cases under strictly controlled environments [48]. Consequently, motion-based approaches are more preferable for 2D-to-3D conversion.

As discussed above, depth can be approximated using motion magnitude, intensity difference, and relative height in a heuristic way. Thus, the problem of how to ensure reasonable accuracy and consistency in estimating depth maps remains unsolved. Again, this refers to two challenges, i.e., dealing with occlusions to recover layered motions, and accurately estimating depth maps. To address these two challenges, a novel method has been proposed to determine depth ordinal and estimate depth maps. Relevant details are presented in the next two sections, respectively.

III. DETERMINING ORDINAL OF DEPTH

Depth ordinal is an important cue for allowing people to identify the depth relationship among the observed objects. For a 2D video sequence with motion information available in the scene, the frame can be separated into a few depth layers regarding its depth discontinuances. Such discontinuities are generally caused by motion inconsistency along the borders of moving objects, i.e., occlusions. As a result, finding depth ordinal has been turned to a new problem on how to determine occlusion in video sequences. Optical-flow based occlusion-reasoning for determining depth ordinal is discussed below.

A. Principles of Optical-flow Based Occlusion Reasoning

Optical flow is a pattern that represents apparent motion of objects, surfaces, and edges in a visual scene caused by relative motion between an observer and the scene [17]. Although the optical flow field appears similar to a dense motion field derived from motion estimation techniques, optical flow is also capable of estimating 3D structure and 3D motion of objects and of the scene itself. While motion estimation focuses on how much each pixel moves from frame to frame, optical flow can also show how pixels move in general. As a result, optical flow has been widely applied in many applications including motion detection, object detection and tracking, dominant image plane extraction as well as visual odometry and robot navigation [17, 30, 36].

The underlying principles of our occlusion reasoning process rely on optical flow to find the motion of each object and

segment a scene to find the occlusion. Let O denote an object (region) in two consecutive frames, I_n and I_{n+1}, and $\Theta(\cdot)$ is defined below to test if O is visible in a frame or not.

$$\Theta(O_n)=\begin{cases}1; & if \quad O \quad visible \quad in \quad I_n \\ 0. & otherwise\end{cases} \tag{2}$$

Using the definitions above, several rules for occlusion reasoning can be summarized as follows.

- **Forward reasoning:** $\Theta(O_n)=1$ and $\Theta(O_{n+1})=0$, i.e., the object is visible in I_n but invisible in I_{n+1}.

  Case 1: O is moving out of I_{n+1} if it is located near the image boundary of I_n and also it is moving towards that boundary.

  Case 2: O is occluded in I_{n+1}, and the occluded region is filled in by other object in front of O.

- **Backward reasoning**: $\Theta(O_{n+1})=1$ and $\Theta(O_n)=0$

  Case 3: O is newly entering I_{n+1} if it is located close to the image boundary.

  Case 4: O was occluded in I_n, and the newly appearing O in I_{n+1} was occluded by the region filled in I_n.

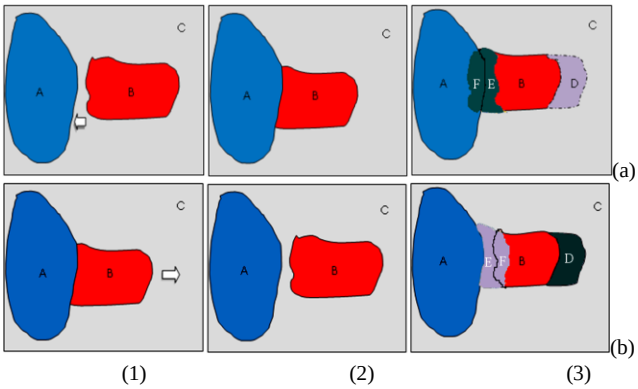


Fig. 3: Illustrations on how occlusion-reasoning works, where (a) and (b) are examples for forward reasoning and backward reasoning, respectively. In (a), a(1) and a(2) refer to two consecutive frames of I_n and I_{n+1}, and a(3) denotes the relationships between occluded and newly appearing regions from I_n to I_{n+1}. In (b), b(1) and b(2) refer to I_{n+1} and I_n, and b(3) denotes the relationships between occluded and newly appearing regions from I_{n+1} and I_n. Different meanings for regions from A to F are discussed in the text.

Considering the above four cases, depth ordinal can be determined as shown in the example shown in Fig. 3, where A, B and C denote three objects/regions in different depth layers. Due to the motion of B between the two frames, occluded and newly appearing regions occur as D, E and F. Using occlusion reasoning, we can analyze the relationships among these objects and/or regions as follows.

Regarding forward reasoning from I_n to I_{n+1}, we can easily find that i) B is in front of C since E in I_n has disappeared in I_{n+1} and it is covered by B in I_{n+1}, and D is appearing in I_{n+1}, where B was in that place in I_n; ii) A is in front of B since F in I_n has disappeared in I_{n+1} whilst it is covered by A in I_{n+1}.

For backward reasoning from I_{n+1} to I_n, we can also find that i) B is in front of C since background region D in I_{n+1} disappears from I_n, and background part E has appeared in I_n from I_{n+1}; and ii) Region F, appears in I_n from I_{n+1}. Since the region to fill this place in I_n is also from object B, so this case is ignored when making decisions.

In summary, we conclude that the depth ordinal among A, B and the background C in Fig. 3 is: A is in front of B; and B is in front of C. In contrast to [20], our occlusion reasoning strategy is much simpler but effective, as it needs neither a Bayesian framework nor object tracking to handle occlusions.

B. Implementation Scheme

Since phase correlation is a fast solution for image matching and insensitive to illumination changes, occlusions, and noise, it is used to estimate optical flow and the technical details can be found in [17]. Two blank matrices, F_{best} and S are introduced to store the best matched flow for each pixel in the frame and the current largest region where this pixel belongs to, respectively. Thus, pixels sharing the same best matched flow will be grouped into the same region, supposing that each region will be best built by its estimated optical flow. If one optical flow assigns a pixel to a larger region than another flow, this optical flow will be

updated as the best matched optical flow for the pixel.

How to apply the four cases for occlusion reasoning using optical flow is explained as follows. For a given frame I_n, we will loop through all its pixels and check whether it has been matched with a pixel in I_{n+1} or not. If the match is found, the pixel has a correspondence in I_{n+1}, i.e., no occlusions due to no motion discontinuity occurring. If unmatched, it applies to case 1 or case 2, where this pixel is either moving out of the scene or occluded in I_{n+1} depending on its location and motion direction in I_n. Similarly, we can also apply such analysis to backward reasoning cases.

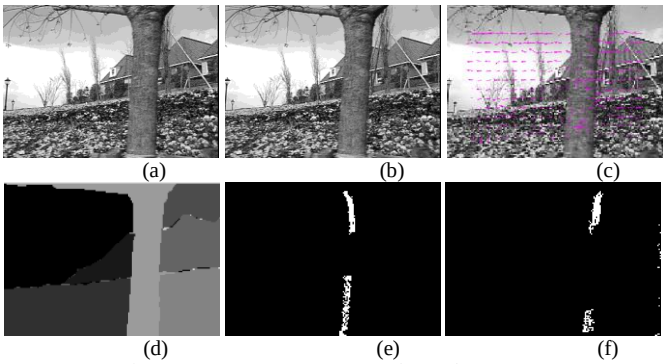


Fig. 4: (a, b) two consecutive frames from 'Flower Garden' sequence; (c) The optical flow of (a); (d) layered representation of objects in (a); (e) the occlusion map from (a) to (b); (f) the occlusion map from (b) to (a).

Using the well-known "flower garden" sequence, Fig. 4 shows results of estimated depth ordinal using optical flow analysis and our occlusion-based reasoning, where (a) and (b) represent two consecutive frames, and (c) is the optical flow estimated from (a) to (b). Results of estimated depth ordinal are illustrated in (d), where (e) and (f) show occlusion maps from (a) to (b) and (b) to (a), respectively. According to the distance between the objects to the camera, the tree, flowers and the house appear in different depth layers. This can be clearly found in our estimated depth ordinal in Fig. 4(d), where the brighter regions are closer to the camera than darker ones. Since the objects are stationary with the camera moving from left to right in the video, the appearance and directions of determined optical flow and occlusion maps are justified.

IV. Object-based Disparity Estimation

In this section, we discuss how consistent depth information is estimated, using an object-based strategy. Three key stages are discussed in detail including layered object extraction, matching with a depth library, and depth refinement.

A. Extracting Layered Objects

Unlike conventional methods in which depth information is recovered on the basis of pixels [12, 14, 18, 37, 47], edge/contours [25-26, 52], or blocks/regions [27, 31, 42, 50, 52], in our system object-based approach is adopted as it is consistent with our human perception [18, 20, 22]. Thus, object-based depth estimation has now attracted increasing attention and some of the recent work is summarized below.

In [20], an object-based method is proposed for stereoscopic image generation, using bi-directional 2D motion estimation for the recovery of rigid motion and structure with a Bayesian framework adopted to handle occlusions. The object mask is estimated through tracking using extended Kalman filtering (EKF) and Kanade-Lucas-Tomasi feature tracker (KLT). An initial segmentation is achieved using K-means, and depth information is decided using shape from motion and estimated focus length by EKF with arbitrarily selected initial values. In [29], an unsupervised object segmentation scheme is proposed, which requires user interaction to assign depth for intra-objects. The segmentation process itself uses anisotropic filtering applied on the difference image between the original frame and the estimated background frame. However, the test scene is relatively simple as there is only one foreground object with no occlusions. In [18] and [22], object-based 3D video generation is presented, using objects and depth ordinal defined in a MPEG-4 sequence, hence object based segmentation is skipped in these systems.

In our system, segmentation of objects is achieved in two steps. First, for each image the determined depth ordinal is considered as an initial segmentation. Second, the mask of each depth layer is extracted and input to our improved seeded region-growing method for refinement. For a given depth layer in image I_n, let Ω be its corresponding binary mask and Λ be the remaining mask in I_n excluding Ω, i.e., $\Lambda = I_n - \Omega$, the refinement process is given as follows.

1. The skeletons of Ω and Λ are extracted and denoted as S_o and S_b and used as a group of seeds for growing of foreground and background regions, respectively. Also we denote R_o and R_b as foreground and background regions under growing, where initially we have $R_o = S_o$ and $R_b = S_b$.

2. In each loop, all the outer neighboring boundary pixels of R_o and R_b are obtained as C_o and C_b. The minimum distance between pixels in C_o and C_b to R_o and R_b are then determined as d_o and d_b below, where $\overline{R}_o$ and $\overline{R}_b$ refer to the average intensity of pixels in R_o and R_b, respectively.

$$d_o = \arg\min_{c_k \in C_o} \left| c_k - \overline{R}_o \right| \tag{3}$$

$$d_b = \arg\min_{c_k \in C_b} \left| c_k - \overline{R}_b \right| \tag{4}$$

3. If $d_o < d_b$, the pixel in C_o which has the minimum distance to R_o is grown into R_o. If $d_b < d_o$, the pixel in C_b which has the minimum distance to R_b is grown into R_b.
4. With updated R_o or R_b, go to step 2 for another loop until no change of R_o or R_b is made.
5. Each of the remaining pixels is checked and merged into R_o or R_b if the majority of its neighboring pixels belong to the corresponding region.

To further improve the robustness of region growing, unreliable skeleton pixels in S_o and S_b are filtered if they are located close to the boundary of Ω and Λ. Further details as well as evaluative results in comparison with conventional approaches can be found in [57].

B. Depth/Disparity Estimation

After having extracted objects from the different depth layers, the next step is to estimate the corresponding depth maps. As 2D videos do not normally have sufficient true depth information for stereo conversion, it becomes extremely difficult to extract their disparity values to reconstruct their stereo version. In our paper, a hybrid solution is adopted to ensure the consistency of estimated depth maps, and details of which are discussed as follows.

1) Recovery of depth maps

Let O denote an extracted object and Φ be a library of objects with corresponding stereo-pair images. Therefore, the disparity for each object in Φ can be accurately determined using stereo constraints. If O is successfully matched with one of the images in Φ, its disparity can be recovered using the disparity attached with that image. Since the disparity from the stereo pair is accurately estimated, the accuracy and consistency of the depth map estimated for O are ensured.

To enable effective matching of objects, the concept of content-based image retrieval is borrowed where textural representation using local binary pattern (LBP) is employed [24]. For a given pixel x, the associated LBP value $b(x)$ is determined using its eight neighboring pixels w_i as follows:

$$b_x = \sum_{i \in [0,7]} 2^i b_i(x) \tag{5}$$

$$b_i = \begin{cases} 1 & if \ x \geq w_i \\ 0 & otherwise \end{cases} \tag{6}$$

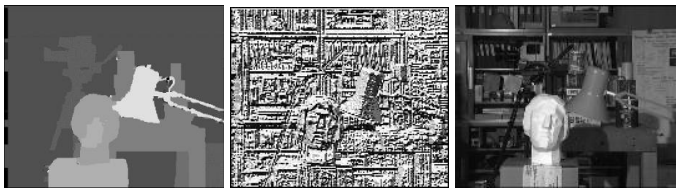


Fig. 5: One example to show relationships among disparity maps (left) and texture maps (middle) via stereo pair using the left-image of 'tsukuba' (right).

For an image or object region I, we can determine the corresponding LBP values and form another image as texture maps of I. For the well-known stereo pair images 'tsukuba', Fig. 5 shows the real disparity/depth maps and the texture maps determined using LBP. As it can be seen, each disparity region can correspond to a region of similar texture maps. This verifies the effectiveness of the proposed approach in matching-based depth recovery.

For each determined texture map, its histogram can be attained by counting the occurrence of each possible LBP value, which is certainly within the range of [0, 255]. This histogram is then used as a key to characterize textural feature of the image/region. Let H_{2D} be the LBP histogram of O and H_i be the histogram of objects in Φ, the best matched histogram H_{best}

is defined as follows, where $\|.\|$ refers to Euclidean norm, and M is the number of objects in Φ.

$$H_{best} = \arg\min_{i\in[1,M]}\left(\|H_{2D} - H_i\|\right) \tag{7}$$

Then, the corresponding image to H_{best} and its associated disparity maps are attained as Φ_{best} and Γ_{best}. As the size and spatial position of Φ_{best} may be different from the one in the 2D video, the adjusted depth D_o is obtained as

$$D_o = \frac{H(\Phi_{best})/T(\Phi_{best})}{H(\mathrm{O})/T(\mathrm{O})}\Gamma_{best} \tag{8}$$

where $H(\cdot)$ and $T(\cdot)$ refer to functions that determine the height and the vertical resolution of the object (image), respectively. When the height is divided by the resolution, an actual measurement of height is attained. Hence Eq. (8) is using relative height to adjust the estimated depth, and a similar strategy can also be found in [52].

By contrast, if a satisfactory matching cannot be retrieved for O, i.e., O is excluded from the disparity library Φ, and a motion-based approach is employed to estimate the disparity maps using the average of horizontal motion magnitude, where N_o is the number of available optical flows in O, and $v_i^{(x)}$ refers to the horizontal part of the i^{th} displacement vector.

$$D_o = N_o^{-1}\sum_i |\, v_i^{(x)}\,| \tag{9}$$

2) Ordinal-based depth regularization

Although depth maps have been extracted for the objects, their values have to be regulated in order to be consistent with the depth ordinal. To achieve this, a linear weighting based regularization process is employed. Assume we have L layers of depth determined, and the average depth for each layer is denoted as $\overline{D}_l \,|\, l\in[1,L]$, where $l=1$ is the closest layer to the camera. A summed average depth D_{sum} is attained as

$$D_{sum} = \sum_{l=1}^{L}\overline{D}_l \tag{10}$$

The regulated depth for each layer is then determined below, which is further applied for stereo video generation.

$$D_l^{'} = \mathrm{K}D_{sum}^{-1}\sum_{i=l}^{L}D_i\,, \qquad l\in[1,L] \tag{11}$$

where K is a constant for normalization. According to the input image in Fig. 4(a), the estimated depth maps are given as the left-image in Fig. 6. Please note that this regularization process has greatly improved the consistency of the estimated depth, and the relevant results are reported in Section V.

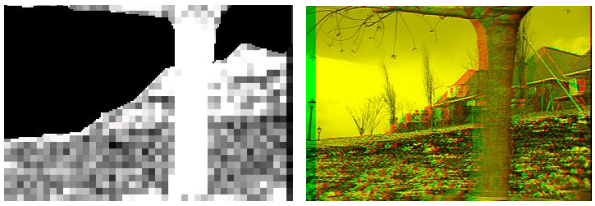


Fig. 6: Extracted depth (disparity) maps (left) and generated stereoscopic frame (right) for Fig. 4(a).

C. Stereo Visualization of Estimated Depth Maps

To enable stereo viewing and visual assessment of the stereo effect of estimated depth maps, stereoscopic videos are generated by taking each original video frame as the left or reference frame, and constructing a right frame using the predicted disparity values. Since two neighboring object regions in different depth layers contain disparity discontinuity, it can result in holes in the newly created right frame if occlusion is detected. To solve this problem, linear interpolation is applied to fill these holes. Afterwards, these two frames are combined together into an anaglyph by incorporating the left and right frames into the red and green component of a synthesized image, respectively. Consequently, the aimed 3D effect from the final converted stereoscopic image can be viewed using spectacles with red and green filters.

For the original image in Fig. 4(a), the generated stereo image is shown in Fig. 6. To visually inspect the stereo effect of the newly generated frame, we need to wear red-green or red-blue paper glasses. Other display options are also available by using digital micro-mirror device and high-contrast shutter glasses [39], anaglyph images with stereo glasses [29], and virtual image(s) rendered using DIBR techniques [11, 40, 41, 43]. However, our display method is the cheapest and the most efficient option suitable for academic research laboratories, where faculties are often under-funded.

V. EXPERIMENTAL RESULTS

In this section, we design two experimental phases to evaluate the performance of the proposed algorithm. The first phase is designed to evaluate the accuracy of the proposed object segmentation, and the second phase is to evaluate the consistency of estimated depth maps. Quantitative and qualitative criteria are applied to the two phases of experiments, respectively.

In our experiments, a total of eight video sequences have been used for evaluation including seven publicly available ones ("Flower garden", "Hall monitor", "Miss America", "Mountains", "Silence", "Interview" and "Orbi") and one of our own sequences "Aisle1". These cover a wide range of videos with various levels of complexity. The relatively simple sequences are those with a fixed background such as "Miss America" and "Silence", where only one foreground object is defined. Sequences of middle complexity are "Hall monitor", "Aisle1", and "Interview", in which motion caused by objects or camera are involved. The most complex scenes can be found in sequences "Mountains", "Flower garden", and "Orbi", in which perspective views with multiple naturally textured objects is contained. Except for the results from "Flower garden" sequence given before, results using other sequences are shown in Fig. 7 and Fig. 8 and further discussed as follows.

A. Performance of Object Segmentation

For objective evaluation of the object segmentation scheme, ground truth of layered objects is manually extracted. By comparing segmented results with the ground truth, the percentage of correctly segmented pixels is attained to measure the accuracy in segmenting objects. The average accuracy over the whole sequence is then obtained to evaluate the effectiveness of our segmentation method. For the six test sequences, the average accuracy values achieved are listed in Table 1, and the segmented results are also shown in Fig. 7.

Table 1: Accuracy of object segmentation in six sequences

Videos / **Results**	*Miss America*	*Silence*	*Aisle1*	*Hall Monitor*	*Flower garden*	*Mountains*
Complexity	low	low	medium	medium	high	high
# objects	2	2	5	2-3	4	6
Accuracy	92.6%	93.3%	84.2%	85.7%	80.2%	81.5%

For statistical consistency in determining the correct rate, the background layer is also considered as an object in each sequence. In Table 1, the complexity of the sequence and the number of objects in the layered scene representation are also given. As it can be seen, this number can be considered as a good measure of the relative complexity of the scene. For scenes of low complexity with no more than two objects, a quite high average accuracy of about 93% is achieved. For sequences of medium complexity, this accuracy drops to around 85%. While for sequences of high complexity, the accuracy achieved is only about 80%.

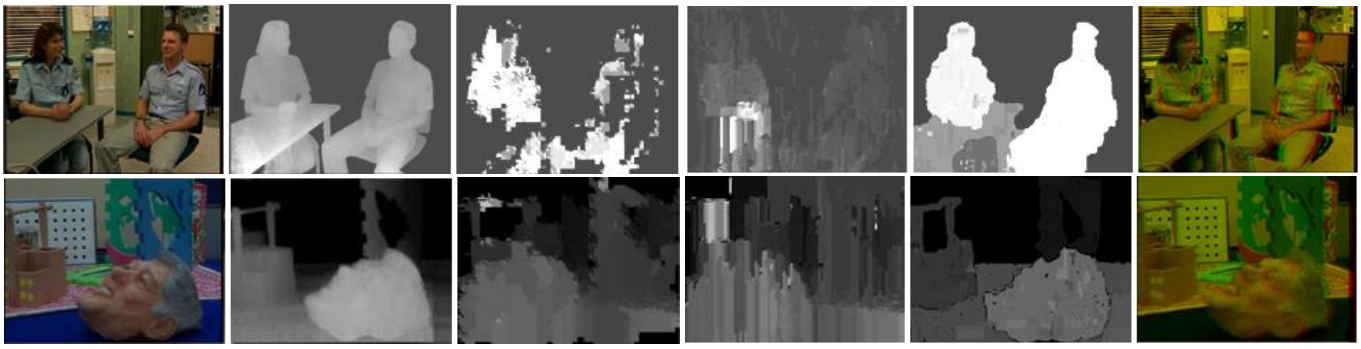
Fig. 7: Results extracted from five test sequences including "Miss America", "silence", "Aisle1", "Hall monitor", and "Mountain" (from left to right). From top to bottom, four rows respectively correspond to original image, segmented results, estimated depth maps, and synthesized stereo image.

As shown in Fig. 7, the reasons for degraded accuracy in sequences of high and medium complexity can be explained as follows. Firstly, limited contrast makes objects hard to identify, such as the small box in "Aisle1" and far-away views in "Mountains" and "Flower garden" sequences. The second reason is that a change of illumination, such as with unstable lighting conditions and shadows, may lead to incorrectly segmented objects as found in "Aisle1" and "Hall monitor" sequences. Further studies are required for developing solutions for object segmentation which are more accurate and robust.

B. Performance of Estimated Depth Maps

To evaluate the effectiveness of our estimated depth maps, subjective and qualitative criteria are employed, where a group of users are invited to watch and grade the quality of generated stereo videos. Before the formal evaluation, training is given to the group users using true stereo video sequence to help them gain a better understanding of the stereoscopic feeling. Then, users are asked to give a satisfaction score of the stereo feeling for each generated video. The score is from 1 to 10, where 1 stands for no stereoscopic feeling and 10 for strongest stereoscopic feeling. Finally, the average score is obtained and used as a measure of the quality of the generated stereoscopic videos.

In total, results from three solutions are compared. In the first solution, depth/disparity maps are estimated using motion only cues as defined in Eq. (9); whilst the second solution uses the hybrid solution defined in Eq. (8). The third solution is the addition of depth ordinal based regularization to the second solution. The performances of these three solutions were evaluated using six test sequences and the results are reported in Table 2. From the table several observations can be made. Firstly, Solution 1 works if fast motion exists in the scene, e.g., the "Flower garden" sequence. Otherwise, it may generate quite poor results. Secondly, Solution 2 works better for sequences with a still background and single object motion, such as "Silence" and "Miss America". Otherwise, it may generate worse results than solution 1. However, the overall accuracy of solution 2 is much better than that of solution 1, which means that introducing texture based image matching indeed can improve the accuracy of estimated depth maps. Thirdly, Solution 3 yields the best results among the three solutions. This finding validates the effectiveness and robustness of the proposed depth-ordinal regularization scheme.



(a) Original sequence (b) ground truth of depth (c) depth from [18] (d) depth from [34] (e) depth of ours (f) reconstructed image

Fig. 8: Evaluation of estimated depth using sequences of "Interview" (top) and "Orbi" (bottom), where our results (e-f) are compared with the ground truth (b) and results from two state-of-art algorithms in [18] and [34] (c-d).

To further evaluate the accuracy of the estimated depth maps, our method was compared with two state-of-the-art methods which were proposed in [18] and [34], using two sequences "Interview" and "Orbi" with available ground truth of depth maps. As seen from the results shown in Fig. 8, consistent depth measurements are achieved for objects in the sequences with our object-based processing, whilst results from [18] and [34] fail to maintain such consistency. In addition, apparent and reliable depth ordinals can be found in our results when these are compared with the ground truth. This confirms the effectiveness of our occlusion reasoning scheme in determining depth ordinal. In contrast, depth ordinal is hardly observable in the results from the two benchmarking methods. Finally, it is worth noting that there exists depth discontinuity in our estimated results. To overcome this drawback, adaptive smoothing filtering [41] and asymmetric filtering [11] may be applied as post-processing to improve the smoothness of the depth map.

Table 2: Scoring of estimated depth maps under 3 solutions

Solutions / **Videos**	*Solution 1*	*Solution 2*	*Solution 3*
	Score(cent)	*Score(cent)*	*Score(cent)*
Miss America	27 (45.0%)	46 (76.7%)	49 (81.7%)
Silence	26 (43.3%)	47 (78.3%)	51 (85.0%)
Aisle1	37 (61.7%)	31 (51.7%)	36 (60.0%)
Hall monitor	36 (60.0%)	35 (58.3%)	42 (70.0%)
Flower garden	46 (76.7%)	40 (66.7%)	51 (85.0%)
Mountains	40 (66.7%)	38 (63.3%)	44 (73.3%)
Average	35.3 (58.9%)	39.5 (65.8%)	45.5 (75.8%)

C. Computational Complexity

Although it is hard to accurately analyze the complexity of the proposed approach, processing time used by each core component of our method is considered a relative indicator for this purpose. For the "Interview" sequence with a resolution of 720×756, on average, it takes 4.66 seconds to generate one 3D image, in which 80.29%, 7.98% and 8.19% of the time are spent for calculating optical flow, segmenting layered objects and querying for disparity maps, leaving 3.54% for other tasks like occlusion reasoning, depth regularization etc. As can be seen, optical flow determination is the most time-consuming part in our system. To approximate optical flow using existing motion vectors within compressed videos might be useful to significantly improve the overall efficiency of the system [58].

VI. CONCLUSIONS

A novel 2D-to-3D video conversion method is proposed, aiming at effective stereoscopic content generation for 3D-TV applications. Two main problems are addressed, i.e., how to determine depth ordinal and how to consistently estimate depth maps from monoscopic sequences. Key contributions of the paper can be summarized as follows.

First, a straightforward occlusion-reasoning scheme is proposed to determine depth ordinal. Using bi-directional optical flows, the relationships among matched or unmatched regions are analyzed to identify possibly occluded regions. As a result, layered representation of objects is attained and used for the determination of depth ordinal. Second, a hybrid solution is proposed to determine depth maps in monoscopic sequences. Inspired by the concept of content-based image matching, a library of objects with stereo pair images is used for consistent depth estimation, where LBP based texture feature is employed for matching. When the matching has failed, motion cues are adopted to estimate relevant disparities. More importantly, regularization of estimated depth maps using determined depth ordinal can further improve the consistency of recovered depth. To assess the performance of our algorithm, quantitative and qualitative criteria were adopted for both objective and subjective evaluations. Comprehensive experiments using eight test sequences of various complexity and contents have fully validated the effectiveness of our proposed methodology. Also it has been found that the number of (layered) objects is a good indicator for

measuring the relative complexity of the scene image.

ACKNOWLEDGEMENTS

First, we wish to thank anonymous reviewers for their constructive comments to improve this paper. Second, special thanks to guest editor Dr. J. Tam for providing some of the test data and carefully revising our paper as well as Dr C. Watkins and Dr. G. Gonzalezcastro of Bradford University for proofreading this paper.

REFERENCES

[1] C. Fehn, P. Kauff, etc., "An evolutionary and optimized approach on 3D-TV," *in Proc. Int. Broadcasting Conf.*, pp. 357-365, 2002

[2] IMAX, http://en.wikipedia.org/wiki/IMAX

[3] L. K. J. Meesters, W. A. Ijsselsteijn, and P. J. H. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Trans. Circuits Systems for Video Technology*., vol. 14, no. 3, pp. 381-391, March 2004

[4] S. Barnard and T. B. William, "Disparity analysis of images," *IEEE Trans. PAMI*, vol. 2, no.4, pp.333-340, Jul. 1980.

[5] A. Kubota, A. Smolic, M. Magnor, etc., "Multiview imaging and 3DTV," *IEEE Signal Proc. Magaz.*, vol.24, no.6, pp.10-21, Nov. 2007.

[6] R. P. Wildes, "Direct recovery of three-dimensional scene geometry from binocular stereo disparity," *IEEE Trans. PAMI*, vol. 13, no. 8, pp. 761-774, Aug. 1991

[7] T. Kanade, and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Trans. PAMI*, vol.16 no.9, pp.920-932, Sept. 1994.

[8] X. Jiang and M. Lambers, "DIBR-based 3D videos using non video rate range image stream," *in Proc. IEEE Int. Conf. Multimedia and Expo* (ICME), pp.1873-1876, 9-12 July 2006

[9] Y. Taguchi, T. Koike, K. Takahashi, etc, "TransCAIP: a live 3DTV system using a camera array and an integral photography display with interactive control of viewing parameters," *IEEE Trans. Visualization and Computer Graphics*, vol.15, no.5, pp.841-852, 2009

[10] M. B. Kim, J. Nam, W. Baek, etc, "The adaptation of 3D stereoscopic video in MPEG-21 DIA," *Signal Proc.*: *Image Commu*., Vol. 18, no. 8, Special Issue on Multimedia Adaptation, pp. 685-697, Sep. 2003.

[11] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Trans. Broadcasting*, vol. 51, no.2, pp. 191-199, 2005.

[12] Y. L. Murphey, J. Chen, etc., "A real-time depth detection system using monocular vision," *In Proc. SSGRR Conf.*, 2000

[13] G. Wei and G. Hirzinger, "Learning shape from shading by a multilayer network," *IEEE Trans. Neural Networks*, vol. 7, no. 4, 985-995, 1996.

[14] C.-H. Choi, B.-H. Kwon and M.-R. Choi, "A real-time field-sequential stereoscopic image converter," *IEEE Trans. Consumer Electronics*, vol.50, no. 3, pp. 903-910, 2004.

[15] A. Yokota, T. Yoshida, H. Kashiyama, etc., "High-speed sensing system for depth estimation based on depth-from-focus by using smart imager," in Proc. *IEEE Int. Symp. Circuits and Systems*, vol. 1, pp. 564–567, 2005

[16] A.S. Ogale, C. Fermuller, and Y. Aloimonos, "Motion segmentation using occlusions," *IEEE Trans. PAMI*, vol. 27, no.6, pp. 988-992, 2005.

[17] T. Gautama and M. A. V. Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. Neural Networks*, vol. 13, no.5, pp. 1127-1136, Sep 2002.

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

[19] T. Okino, H. Murata, K. Taima, etc., "New television with 2D/3D image conversion technologies," *in Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems III*, vol. 2653, pp. 96-103, 1996.

[20] K. Moustakas, D. Tzovaras, and M. G. Strintzis, "Stereoscopic video generation based on efficient layered structure and motion estimation from a monoscopic image sequence," *IEEE Trans. Circuits and Systems for Video Technology*, vol.15, no.8, pp. 1065- 1073, Aug. 2005

[21] S. H. Lai, C. W. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image," *PAMI*, Vol.14, no.4, pp.405-411, 1992.

[22] M. Kim, S. Park, and Y. Cho, "Object-based stereoscopic conversion of MPEG-4 encoded data," *Lecture Notes in Computer Science*, vol. 3333, pp. 491-498, 2004.

[23] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *IEEE Trans. PAMI*, vol.15, no. 2, pp. 97-108, 1993

[24] G.C. Feng and J. Jiang, "Image extraction in DCT domain," *IEE Proc.: Vision, Image and Signal Proc.*, vol. 150, no. 1, pp. 20-27, 2003.

[25] W. J. Tam, F. Speranza, L. Zhang, etc., "Depth image based rendering for multiview stereoscopic displays: Role of information at object boundaries," in *Proc. SPIE*, vol. 6016, pp. 75-85, 2005

[26] W. J. Tam, A. S. Yee, J. Ferreira, etc., "Stereoscopic image rendering based on depth maps created from blur and edge information," *Stereoscopic Displays & Applications XII*, vol. 5664, pp.104-115, 2005

[27] S. Battiato, A. Capra, S. Curti, etc., "3D stereoscopic image pairs by depth-map generation", *in Proc. 2nd Int. Sympo. 3D Data Proc., Visual. Transmission*, pp. 124-131, 2004.

[28] F. Pitie and A. Kokaram, "Matting with a depth map", in *Proc. IEEE ICIP*, pp. 21-24, 2010.

[29] M. Kunter, S. Knorr, A. Krutz, etc., "Unsupervised object segmentation for 2D to 3D conversion," *in Proc. SPIE* (Stereoscopic Displays and Applications XX), vol. 7237, 7231B, 2009.

[30] D. Sun, S. Roth and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. CVPR*, pp. 2432-2439, 2010.

[31] S. Battiato, S. Curti, M. La Cascia, etc., "Depth map generation by image classification", *In Proc. SPIE, v*ol. 5302, pp. 95-104, 2004

[32] P. Harman, J. Flack, S. Fox, etc., "Rapid 2D to 3D conversion," in *Proc. SPIE*, vol.4660, pp. 78-86, 2002

[33] D. Kim, D. Min, and K. Sohn, "Stereoscopic video generation method using motion analysis," *in Proc. 3DTV Conf.* pp. 1-4, 2007.

[34] I. Ideses, L. P. Yaroslavsky and B. Fishbain, "Real-time 2D to 3D video conversion," *J. Real-Time Image Proc.*, vol. 2, no. 1, pp. 3-9, 2007.

[35] D.C. Burr and J. Ross, "How does binocular delay give information about depth?" *Vision Research*, vol. 19, pp.523-532, 1979.

[36] S. Baker, D. Scharstein, J. P. Lewis, etc., "A database and evaluation methodology for optical flow," Microsoft Research Report (TR-2009-179), http://research.microsoft.com/pubs/117766/ofevaltr2.pdf, 2009.

[37] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," *IEEE Trans. Broadcasting*, vol.54, no.2, pp.188-197, June 2008.

[38] Y. Feng, J. Jiang, and S. S. Ipson, "A shape-match based algorithm for pseudo-3D conversion of 2D videos," *in Proc. ICIP*, pp.808-811, 2005.

[39] D. C. Hutshison and H. W. Neal, "The design and implementation of a stereoscopic microdisplay television," *IEEE Trans. Consumer Electronics*, vol. 54, no. 2, pp. 254-261, May 2008.

[40] Y. Feng, D. Li, K. Luo, and M. Zhang; "Asymmetric bidirectional view synthesis for free viewpoint and three-dimensional video," *IEEE Trans, Consumer Electronics*, vol.55, no.4, pp.2349-2355, November 2009

[41] Y. K. Park, K. Jung, Y. Oh, etc., "Depth-image-based rendering for 3DTV service over T-DMB," *Signal Proc.: Image Communication*., vol.24, issues 1-2, pp. 122-136, January 2009.

[42] S.-Y. Kim, E.-K. Lee, and Y.-S. Ho, "Generation of ROI enhanced depth maps using stereoscopic cameras and a depth camera," *IEEE Trans. Broadcasting*, vol.54, no.4, pp.732-740, Dec. 2008

[43] S.Y. Kim, S.B. Lee, and Y.S. Ho, "Three-dimensional natural video system based on layered representation of depth maps," *IEEE Trans. Consumer Electronics*, vol.52, no.3, pp.1035-1042, Aug. 2006

[44] G. Wang, and Q.M.J. Wu, "Perspective 3-D Euclidean reconstruction with varying camera parameters," *IEEE Trans. Circuits and Systems for Video Technology*, vol.19, no.12, pp.1793-1803, Dec. 2009

[45] Y. Feng, J. Jayaseelan, and J. Jiang, "Cue based disparity estimation for possible 2D to 3D video conversion," in *Proc. Int. Conf. Visual Information Engineering* (VIE), 2006.

[46] Y. Feng and J. Jiang, "Pseudo-stereo conversion from 2D video", in *Proc. Int. Conf. Advanced Concepts for Intelligent Vision Systems*, Lecture Notes in Computer Science, vol. 3708, pp. 268-275, Oct 2005.

[47] L. Zhang, B. Lawrence, D. Wang, etc., "Comparison study on feature matching and block matching for automatic 2D to 3D video conversion," *in Proc. CVMP*, pp. 122-129, 2005.

[48] E. Stoykova, A. Alatan, P. Benzie, etc, "3-D time-varying scene capture technologies-a survey," *IEEE Trans, Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1568-1586, Nov. 2007.

[49] W. J. Tam and L. Zhang, "3D-TV content generation: 2D-to-3D conversion," *in Proc. ICME*, pp. 1869-1872, 2006

[50] X. L. Deng, X.H Jiang, Q. G. Liu, etc., "Automatic depth map estimation of monocular indoor environments," *in Proc. Int. Conf. on MMIT*, pp.646-649, 2008

[51] M. Fieseler and X. Jiang, "Registration of depth and video data in depth image based rendering," *in Proc. 3rd 3DTV-Conference*, 2009.

[52] Y. J. Jung, A. Baik, J. Kim, etc, "A novel 2D-to-3D conversion technique based on relative height depth cue," *in Proc. SPIE*, vol. 7237, 2009.

[53] J. Kim, A. Baik, Y. J. Jung, etc, "2D-to-3D conversion by using visual attention analysis," *in Proc. SPIE*, vol. 7524, 2010.

[54] W.J. Tam, F. Speranza, C. Vázquez, L. Zhang, "Temporal sub-sampling of depth maps in depth image based rendering of stereoscopic image sequences," *in Proc. SPIE*, vol. 7237, 2009

[55] W.J. Tam, C. Vázquez, and F. Speranza, "Three-dimensional TV: A novel method for generating surrogate depth maps using colour information," *in Proc. SPIE*, vol. 7237, 2009.

[56] A. Smolic, K. Mueller, P. Merkle, etc. "3D video and free viewpoint video - technologies, applications and MPEG standards," *in Proc. ICME*, pp.2161-2164, 2006.

[57] Y. Feng, H. Fang, and J. Jiang, "Region growing with automatic seeding for semantic video object segmentation," in *Lecture Notes in Computer Science*, vol. 3687, pp. 542-549, 2005

[58] M. T. Coimbra and M. Davies, "Approximating optical flow within the MPEG-2 compressed domain," IEEE Trans. Circuits and Systems for Video Technology, vol. 15, no. 1, pp. 103-107, 2005.