# Wavelet based stereo images reconstruction using depth images

Ljubomir Jovanov, Aleksandra Pižurica and Wilfried Philips

Telecommunications and Information Processing Department, Ghent University, Sint Pietersnieuwstraat 41, 9000 Ghent, Belgium;

## ABSTRACT

It is believed by many that three-dimensional (3D) television will be the next logical development toward a more natural and vivid home entertaiment experience. While classical 3D approach requires the transmission of two video streams, one for each view, 3D TV systems based on depth image rendering (DIBR) require a single stream of monoscopic images and a second stream of associated images usually termed depth images or depth maps, that contain per-pixel depth information. Depth map is a two-dimensional function that contains information about distance from camera to a certain point of the object as a function of the image coordinates. By using this depth information and the original image it is possible to reconstruct a virtual image of a nearby viewpoint by projecting the pixels of available image to their locations in 3D space and finding their position in the desired view plane. One of the most significant advantages of the DIBR is that depth maps can be coded more efficiently than two streams corresponding to left and right view of the scene, thereby reducing the bandwidth required for transmission, which makes it possible to reuse existing transmission channels for the transmission of 3D TV. This technique can also be applied for other 3D technologies such as multimedia systems.

In this paper we propose an advanced wavelet domain scheme for the reconstruction of stereoscopic images, which solves some of the shortcommings of the existing methods discussed above. We perform the wavelet transform of both the luminance and depth images in order to obtain significant geometric features, which enable more sensible reconstruction of the virtual view. Motion estimation employed in our approach uses Markov random field smoothness prior for regularization of the estimated motion field.

The evaluation of the proposed reconstruction method is done on two video sequences which are typically used for comparison of stereo reconstruction algorithms. The results demonstrate advantages of the proposed approach with respect to the state-of-the-art methods, in terms of both objective and subjective performance measures.

**Keywords:** depth images, wavelets, stereo reconstruction

## 1. INTRODUCTION

First attempts to generate stereoscopic images were made in 1838 by Sir Charles Wheatstone, a british scientist and inventor, who developed a device based on a system of mirrors, which enabled the viewers to have three dimensional impression. Next significant step in the development of three-dimensional techniques was in 1903, when brothers Lumière showed their first three dimensional movies. The first full-lenght three dimensional movie was made in 1922. As opposed to short movies made by Lumière brothers, this movie could be watched simultaneosly by multiple viewers. Three dimensional techniques were first introduced in television in 1928. by one of the pioneers of television, John Logie Baird. In the begining of the fifth decade of the last century a large number of three-dimensional movies were made. Despite the initial success, the interest for three-dimensional movies started to decrease mainly because of unsuficient experience of movie makers, low quality of movies and unexperienced operators in the cinemas. Broadcast of three-dimensional TV programs had a similar destiny, in a sense that first commercial systems were launched thirty years

Further author information: (Send correspondence to Ljubomir Jovanov)
Ljubomir Jovanov: E-mail: ljj@telin.ugent.be, Telephone: +32 9 264 34 12, A. Pižurica is a postdoctoral researcher of FWO Flanders, Belgium

after first broadcast occured. Because of the poor quality of these analogue services, interest was rather low, and the broadcast was stopped. New possibilities for three-dimensional TV broadcast arised with the introduction of digital TV services. Numerous European-funded projects were launched in the early 1990's to develop technologies needed for digital transmission of 3D images.

As a consequence of the new interest in three-dimensional broadcast services, MPEG2 compression standard was extended with the techniques for compression of stereo video sequences. The part of the standard related to stereo video coding was named Multiview Profile. In this part of the standard, left-eye view is encoded in compliancy with the MPEG2 main profile, to provide backward compatibility with existing digital TV receivers, while the right-eye view is encoded as an enhancement layer. Despite wide acceptancy of the MPEG2 as a standard for a digital TV broadcast, commercial 3D broadcast systems based on MPEG2 haven't been offered so far. In Japan, a system was made that integrates high definition TV and 3D transmission [1]. Currently the most important develpoments for 3D broadcast systems are inititated by major Japanese electronics companies and MPEG group.

Initial proposals for transmission of 3D images were based on simulataneous broadcast of left and right stereoscopic view as in [2]. That approach requires high bandwidth for transmission. The most recent proposals for 3D broadcast systems are based on transmission of monoscopic color video and per-pixel associated depth information. This representation enables creating one or more virtual views from monoscopic view and appropriate depth map. Depth image based rendering enables creating close views of the scene, using information from the depth maps, as if they were captured with another camera from a different viewpoint.

However numerous problems in DIBR are still limitting the quality of the reconstructed images. The main problem is how to deal with the newly exposed areas (holes) in virtual images. These appear due to disocclusion of regions of objects that are visible only from the new viewpoint, which is generated from an existing view. The simplest way to solve this problem is to replace the missing pixels with the average of non-occluded neighoboring pixels or more complex interpolation polynomials [3]. Interpolation techniques above mentoned are known to produce visible artefacts, whose level depends on scene geometry.

Several algorithms have been suggested in order to reduce these interpolation artefacts. The layered-depth-image (LDI) method, proposed in [4], uses multiple images of the original scene, taken from multiple angles and their corresponding depth maps. Compared to other existing methods this approach offers quite accurate image reconstrucion, but at the expense of a relatively high transmission bandwidth and computational cost. Alternative approaches involve pre-processing of the depth maps [5]. Using this approach, percentage of removed disoclussion artifacts increases with the strenght of smoothing of depth maps [6], [7] and [3]. In a recently described technique [5], the authors use asymmetrical filtering to avoid geometry distortions present in prevoius techniques [6], [7] and [3].

Small unfilled regions such as single lines and isolated pixels are usually produced by noise or deinterlacing artefacts. Large unfilled regions are considered as occluded scene caused by perspective or motion. None of the approaches above mentioned use information about the motion in the video/depth sequence. The existing methods also do not take into account the geometry of the scene, and make interpolation by simple filtering. Although previous methods remove most of the artefacts, this is not done in an optimal way, and annoying artefacts are introduced in ocluded regions. Finally, these methods do not take into account possible distortions during transmission such as noise or blocking and deinterlacing artefacts.

In this paper we propose an advanced wavelet domain scheme for the reconstruction of stereoscopic images, which solves some of the shortcommings of the existing methods discussed above. We perform the wavelet transform of both the luminance and depth images in order to obtain significant geometric features, which enable more sensible reconstruction of the virtual view. Compared to more common pixel-based occlusion correction methods, wavelet-based methods were not reported in the available literature. Wavelets provide an efficient means for approximating images with a small number of basis elements, and as such are widely used for image coding. Information about significant details in image is represented with a relatively small number of significant coefficients. These coefficients are interdependent and enables us to reconstruct missing information from existing coefficients.

In the proposed method, scaling coefficients and wavelet coefficients are utilized for motion estimation, both for depth and luminance, within an iterative scheme. The estimated motion fields are dense. Together with the motion vectors, we estimate motion reliability and use it within the proposed reconstruction method. Motion estimation on depth sequences not only helps the motion estimation on the corresponding luminance sequences, but it also enables the detection of some situations (for example zoom) which are otherwise hard to detect. Multiple hypotheses are generated for different resolution levels, which are then optimised within a Bayesian approach using a Markov random field smoothness prior [8]. The proposed motion estimation algorithm starts with block matching with variable blocks size, and then chooses the optimal one. In the case when the motion vectors are reliable, the wavelet coefficients belonging to occluded areas are obtained by using the corresponding values from the previous frames. Othervise occluded pixels are calculated by interpolation of the surrounding wavelet coefficients.

In Section 2, we explain virtual views rendering, and in Section 3 we describe motion estimation algorithm used in our paper and novel priors for Markov random field optimization, in Section 4 we describe the algorithm for improving rendered view images. The results are presented in Section 4 and conclusions in Section 5.

## 2. VIRTUAL VIEWS RENDERING

Generation of virtual views or depth-image based rendering is a technique of synthesizing virtual views of a scene using one available monoscopic view and associated depth information. In the first step, images from the monoscopic view are projected into the volume defined by the depth-map. Finally points from the volume are projected on to the plane of the virtual camera, positioned adequately. This process is usualy termed 3D image warping. In the following, we will give the short overview of the process. Let the stereoscopic system consists of two pinhole cameras. If we assume that the coordinate system of the first camera is equal to the world coordinate system, perspective projection equation for the arbitrary 3D point $M = (x, y, z, 1)^T$ can be written as:

$$z_l \mathbf{m_l} = K_\mathbf{l}[I|\mathbf{0}]\mathbf{M} \tag{1}$$

$$z_r \mathbf{m_r} = K_\mathbf{r}[R|\mathbf{t}]\mathbf{M} \tag{2}$$

where $\mathbf{m_l} = (\mathbf{u_l}, \mathbf{v_l}, \mathbf{1})^T$ and $\mathbf{m_r} = (\mathbf{u_r}, \mathbf{v_r}, \mathbf{1})^T$ are projection of the point $M$ on the planes of left and right camera, $R$ is $3 \times 3$ matrix and $t$ $3 \times 1$ vector which define the rotation and translation respectively, from the world coordinate system into the camera coordinate system and $u$ and $v$ are the coordinates in the projection planes. $K_l$ and $K_r$ are upper triangular matrices which specify the internal parameters of the two cameras and $z_l$ and $z_r$ are the depths of the scene in the coordinate system of the left and right camera, respectively. By substituting (1) into (2) we obtain the affine disparity equation, which defines the correspondence between pixels of images from two views of the scene:

$$z_r \mathbf{m_r} = z_\mathbf{l} K_\mathbf{r} R K_\mathbf{l}^{\mathbf{-1}} \mathbf{m_l} + K_\mathbf{r}\mathbf{t} \tag{3}$$

This equation can be used for creating an arbitrary novel view from a known reference view. Required input parameters for generating virtual view are vector $t$ which defines the translation of the virtual view, matrix $R$ which defines the rotation of the virtual view and parameters of the virtual camera, which are defined by matrix $K$. If the depth information is available for each point of the rigid body, it is possible to generate an arbitrary virtual view.

The above principle can also be used for generation of pairs of virtual views, that correspond to left and right view. The first step in this process is the selection of the convergence distance $Z_c$ or *zero-parallax setting* (ZPS). There are three approaches to accomplish this. The first one is so-called *toed-in* method, where the zero-parallax setting is determined by joint inward rotation of the right and left-eye view cameras, which is shown in Fig. 2. The second approach is so-called *shift-sensor* approach, where the convergence

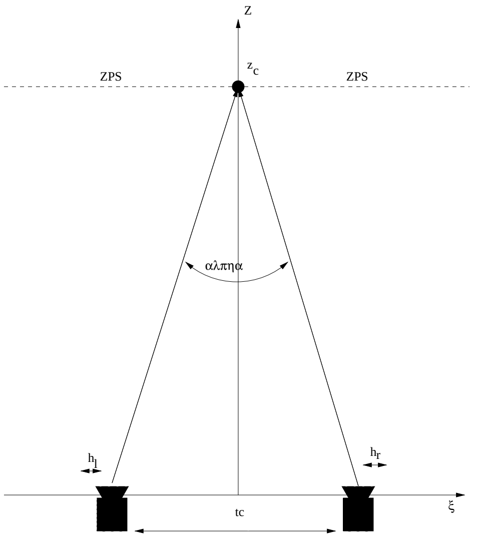

**Figure 1.** Shift sensor approach for determining the zero parallax setting

distance is determined by a small shift of the parallel positioned CCD sensors, as shown in Fig. 1. In the third method, which is the simplest one, the zero parallax setting is chosen by shifting the depth map. Here the convergence distance is chosen as: $Z_c = \frac{Z_{near}-Z_{far}}{2}$, where $Z_{near}$ and $Z_{far}$ are the nearest and the farthest clipping plane of the depth map. The next step in the algorithm is normalization of the depth values to interval of $[-0.5, 0, 5]$.

In this paper, we consider only the simplest and at the same time most commonly used parallel camera configuration for generating virtual stereoscopic images using a central image and the associated depth map. Parallel camera configuration for generating stereoscopic images, and relevant parameters for generating images are shown in Fig. 3, where $c_c$ denotes the viewpoint of the original center image, $c_l$ and $c_r$ are the viewpoints of the virtual generated left and right eye images, $f$ is the focal distance of three cameras and $t_x$ is the baseline distance between virtual cameras. Under this geometry constraint, vertical coordinates of any 3D point projected on left and right image plain remain the same. Thus the point $p$ with the coordinates $(X, Y, Z)$ are projected onto the image planes of the three cameras $(x_l, y)$, $(x_c, y)$ and $(x_r, y)$. It can be shown that the $x$ coordinates of the left and right projections from Fig. 3 can be obtained as [3]:

$$x_l = x_c + \frac{t_x}{2}\frac{f}{Z}x_r = x_c - \frac{t_x}{2}\frac{f}{Z} \tag{4}$$

where $x_c$ and $Z$ are taken from the center image and the corresponding depth map. The model parameters, i.e., focal length and baseline distance are chosen in such a way that parallax do not exceed 3

## 3. MOTION ESTIMATION

Motion field estimation is one of the fundamental problems in computer vision. Various solutions to this problem were offered with various success. In some cases, the most important criterion is to reduce complexity of the algorithm at the expense of the higher compensation error. Motion vectors can be estimated either for the blocks of finite size or for each pixel. Here we introduce stochastic approach for the computation of motion. The approach starts from the well-known concepts from the stochastic modelling and reconstruction, Markov random field models for images and extend them with new energies and cost functions based on wavelets (5) and (6). Motion estimation algorithm used in this paper starts from an initial motion estimate which is obtained on 8x8 blocks. This step speeds-up the estimation of the dense field since it gives good initial estimate of the motion for most of the blocks. Very often motion is estimated

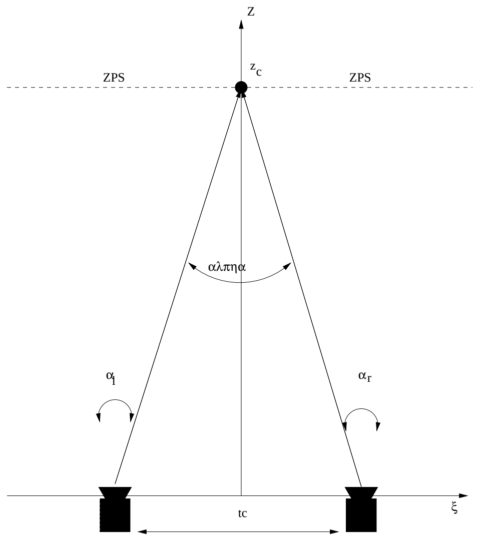

**Figure 2.** Toed in approach for determining the zero parallax setting

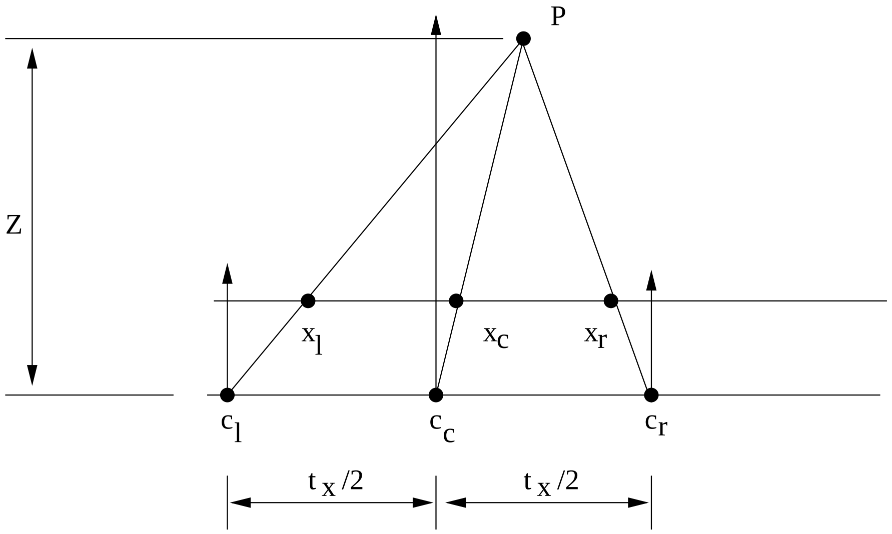

**Figure 3.** Relevant parameters for stereoscopic images generation in parallel camera configuration

by minimizing the mean absolute difference (MAD) of pixels inside blocks. Motion field obtained this way contains large number of wrong estimates, even in the case of full search for each pixel in the block. In this approach, we are using wavelet subbands to estimate motion vectors by minimizing sum of MAD's for each band and MAD of correlation mask which will be defined later. The search process minimizes the following sum:

$$E = \sum_{l=1}^{L}\sum_{o=1}^{3}\sum_{i=1}^{N}\sum_{j=1}^{N} |W_{i,j}^{l,o,k} - W_{i,j}^{l,o,k-1}| + |CM(i,j)^k - CM(i,j)^{k-1}| \tag{5}$$

where $L$ denotes the number of levels in the wavelet decomposition, $N$ denotes the size of the block, $W_{i,j}^{l,o,k}$ are the wavelet coefficients for level $l$ and orientation $o$ of the wavelet decomposition, and CM is the correlation mask. Correlation mask is defined as sum of the products of the wavelet bands at different levels of decomposition:

$$CM(i,j) = \sum_{o=1}^{3}\prod_{l=1}^{L} |W_{i,j}^{l,o,k}| \tag{6}$$

Correlation mask is a robust edge detector and as such can be used as an important feature for motion estimation. Values of correlation mask for the first frame of the "Interview" sequence is shown in Fig. 4

Initial motion field obtained this way is further optimized in a stochastic manner, by using Markov random fields.

Block based motion estimation techniques do not take into account the object boundaries, which causes large compensation errors near egdes of the moving objects. Motion fields in video sequences consists of a regions of similar vectors, in orientation and lenght, with possible discontinuities at motion boundaries. Here we assume that the motion fields are smooth functions of spatial coordinates, for one time instant, except for some regions where vector intensities and lenghts change abruptly. Motion fields are much smoother than images themselves, due to the fact that motion vector are approximately the same for the whole moving object. Taking these characteristics into account, motion fields can be successfuly modeled using coupled binary and vector Markov random fields (MRF) for motion and motion discontinuity fields $L_t$ and $D_t$.

Let $S$ is a discrete set of $\mathbf{m}$ sites $S = \{1, ..., m\}$ where $1, ..., m$ are indices. The sites in $S$ are related to one another via a neighbourhood system. Neighbourhood system for $S$ is defined as:

$$N = \{N_i | \forall i \in S\} \tag{7}$$

where $N_i$ is the set of sites neighboring $i$. The main properties of the neighboring relationship are that a sites are not neighboring to itself, and that the relationship is mutual. The family of random variables $F$ defined on $S$ is a Markov random field on $S$ with respect to a neighborhood system $N$ if and only if the probabilities of all realizations are strictly greater than zero, and that realization of random field for some specific location depends only on neighboring locations (Markovianity). Properties of scalar MRFs hold for vector MRFS, since the only difference is the definition of the state.

Optimal displacement and line field estimates, for the neigboring frames maximize the joint probability $P(\mathbf{D_t} = \hat{\mathbf{d}}_{\mathbf{t}}, L_{\mathbf{t}} = \hat{l}_{\mathbf{t}} | g_{\mathbf{t}_-}, g_{\mathbf{t}_+})$ of motion and motion discontinuity field given the previous and current frame. From the Bayes rule this probability can be written as:

$$P(\mathbf{D_t} = \hat{\mathbf{d}}_{\mathbf{t}}, L_{\mathbf{t}} = \hat{l}_{\mathbf{t}} | g_{\mathbf{t}_-}, g_{\mathbf{t}_+}) = \frac{P(G_{\mathbf{t}_+} = g_{\mathbf{t}_-} | \mathbf{d_t}, l_{\mathbf{t}}, g_{\mathbf{t}_-}) P(\mathbf{D_t} = \mathbf{d_t}, L_{\mathbf{t}} = l_{\mathbf{t}} | g_{\mathbf{t}_-})}{P(G_{\mathbf{t}+} = g_{\mathbf{t}_+} | g_{\mathbf{t}_-})} \tag{8}$$

Since the expression in numerator does not depend on motion and motion discontinuity fields, it can be neglected when $P(\mathbf{D_t} = \hat{\mathbf{d}}_{\mathbf{t}}, L_{\mathbf{t}} = \hat{l}_{\mathbf{t}} | g_{\mathbf{t}_-}, g_{\mathbf{t}_+})$ is maximized with respect to $(\hat{d_t}, \hat{l_t})$. Taking this into account MAP estimate of the pair $(d_t, l_t)$ can be obtained by maximizing the numerator of Eq. 8:

$$(\hat{\mathbf{d}}_t, \hat{l}_t) = argmax_{(\mathbf{d}_t, l_t)}[P(G_{t_+} = g_{t_-}|\mathbf{d}_t, l_t, g_{t_-})P(\mathbf{D}_t = \mathbf{d}_t, L_t = l_t|g_{t_-})] \tag{9}$$

To maximize all the expressions above, all the constituent probabilities must be known. Using Bayes rule, distribution of the motion model given by $P(G_{t_+} = g_{t_-}|\mathbf{d}_t, l_t, g_{t_-})$ can be written as:

$$P(G_{t_+} = g_{t_-}|\mathbf{d}_t, l_t, g_{t_-}) = P(\mathbf{D}_t = \mathbf{d}_t|l_t, g_{t_-})P(L_t = l_t|g_{t_-}) \tag{10}$$

If $P(\mathbf{D}_t = \mathbf{d}_t|l_t, g_{t_-})$ and $P(L_t = l_t|g_{t_-})$ are Gibbsian then $P(G_{t_+} = g_{t_-}|\mathbf{d}_t, l_t, g_{t_-})$ is also Gibbsian, and the pair $(\mathbf{d}_t, l_t)$ has Markovian properties. Usualy field which depends on image intensities do not affect motion vector model much, first member of previous is considered as independent of $g_{t_-}$. Ussualy motion discontinuities most probably occur at the positions which correspond to object edges, which is modeled by $P(L_t = l_t|G_{t-} = g_{t-})$. Taking the above assumptions into account the probability optimization criterion for displacement field can be modelled with Gibbsian distribution:

$$P(\mathbf{D}_t = \mathbf{d}_t|L_t = l_t) = \frac{1}{Z_d}e^{-\frac{U_d(\mathbf{D}_t=\mathbf{d}_t|L_t=l_t)}{\beta_d}} \tag{11}$$

where $Z_d$ is a normalizing constant, $\beta_d$ is a constant which controls optimization process, or more precise the influence of the neighbouring motion vectors. The energy function is defined as:

$$U_d(\mathbf{D}_t = \mathbf{d}_t|L_t = l_t) = \sum_{c_d=x_i,x_j\in C_d} V_d(\mathbf{d}_t, c_d)[1 - l((x_i, x_j), t)] \tag{12}$$

where $c_d$ is a clique of vectors, while $C_d$ is a set of all such cliques derived from a neighbourhood $N_d$. This function introduces penalties if motion vector has large deviation from neighboring vectors, and line field is not present. When the motion field discontinuity field is present the penalties are not introduced. In this way we perform vector field regularization inside moving objects, while keeping motion discontinuities at the positions where line field is activated. A priori displacement model is introduced through potential function $V_d$ as:

$$V_d(d_t, c_d) = V(d(x_i, t), d(x_j, t)) = ||d(x_i, t) - d(x_j, t)||^2, c_d = x_i, x_j \in C_d \tag{13}$$

where $||.||$ is a norm in $R^2$. For the optimization we use four neighbourhood cliques. Bigger values of parameter $\beta_d$ introduce less constraints on behaviour of motion vectors and vice versa.

Line field is modeled with binary MRF $L_t$ and is described by the Gibbs probability distribution:

$$P(L_t = l_t|G_{t-} = g_{t\_}) = \frac{1}{Z_l}e^{U_l(L_t=l_t|G_{t-}=g_{t\_})/\beta_l} \tag{14}$$

Line energy function $U_l$ is defined as:

$$U_L(L_t = l_t|G_{t-} = g_{t\_}) = \sum_{c_l\in C_l} V_l(l_t, g_{t\_}, c_l) \tag{15}$$

where $c_l$ is a line clique and $C_l$ is a set of all line cliques from the neighbourhood system. Here, eight neighbour cliques are used. One of the novelties introduced in this algorithm is the usage of line potential function which depends on correlation mask defined in (6). A priori probabilities of the line field are conditioned on observations. In our case, potential function is inverse proportional to the values of the correlation mask defined in (5). This means that the penalties on discontinuities are introduced if the line element is present and the corresponding gradient is small. Combining all the above we get the Gibbs aposteriori probability:

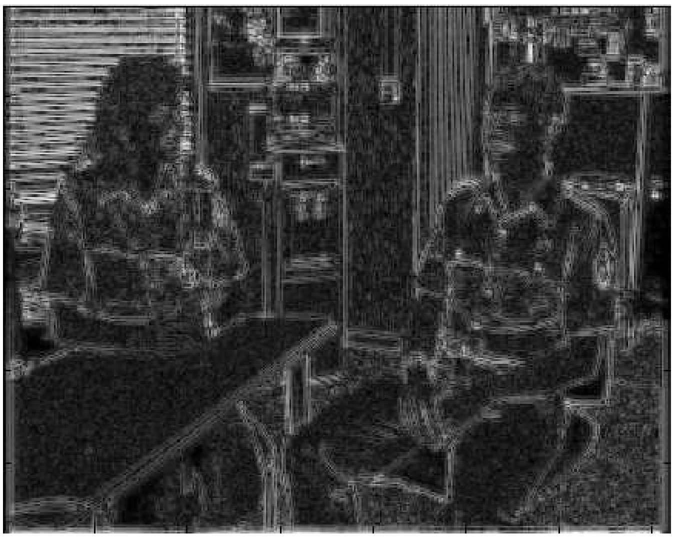

**Figure 4.** Correlation mask for the first frame of "Interview" sequence

$$P(\mathbf{D}_t = \hat{\mathbf{d}}_t, L_t = \hat{l}_t | g_{t_-}, g_{t_+}) = \frac{1}{Z} e^{-U(\hat{\mathbf{d}}_t, \hat{l}_t, g_{t-}, g_{t+})} \tag{16}$$

The data error is modeled as a sum of difference between wavelet coefficients in neighboring frames:

$$r(d(x_i,t), x_i, t) = \sum_{o=1}^{3} \sum_{s=1}^{L} w(x_i - d(x_i,t), t) - w(x_i, t-1) \tag{17}$$

The usage of wavelet coefficients for calculation of the data error is the main novelty in this paper. Wavelets have already been used as a features in numerous block based motion estimation algorithms e.g. [9], [10] and proved ther advantages in terms of compensation error. Corresponding energy function is given as:

$$U_g(g_{t_+} | d_t, g_{t_-}) = \sum_{i=1}^{M_d} [r(d(x_i,t), x_i, t)]^2 \tag{18}$$

Taking all potentials into account we obtain final energy function:

$$U(\hat{\mathbf{d}}_t, \hat{l}_t, g_{t_-}, g_{t_+}) = \lambda_g U_g(g_{t_+} | \hat{d}_t, g_{t_-}) + \lambda_d U_d(\hat{\mathbf{d}}_t | \hat{l}_t) + \lambda_l U_l(\hat{l}_t | g_{t_-}) \tag{19}$$

Sample configurations of motion and motion discontinuities fields are generated through modified Gibbs sampler. Samples for line field are generated in a such a way that the same location is never visited twice in one iteration. Samples for the motion field are generated according to the distribution of motion vectors of the initial estimate of the motion field. In the case of line field configuration is chosen if it decreases the energy of the configuration. Since it is binary field new configurations are generated by switching on or off line elements. In the case of motion field small random variations are added to current motion vectors, according to the distribution of the motion field. Motion vector hypothesis are validated both in depth and luminance in a way that if the estimated motion is translational in both domains and vector directions and intensities have similar values, motion reliability is higher. Changes are accepted if the new configuration reduces the energy of the motion field. Examples of estimated motion and motion discontinuities field are shown in Figures 6 and 5.

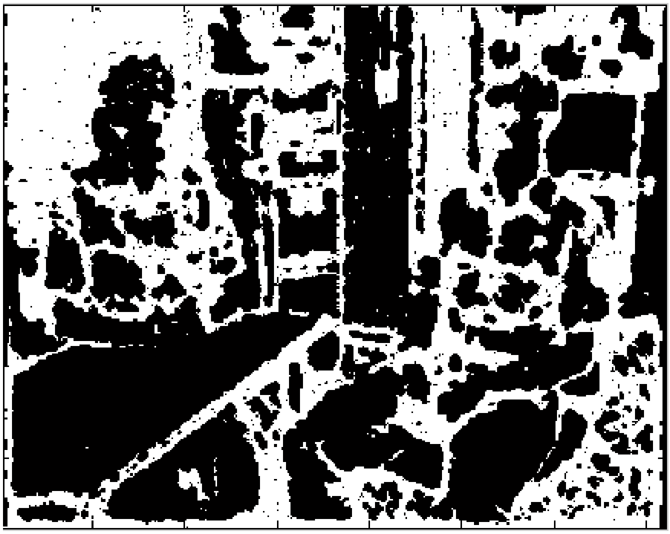

**Figure 5.** Motion discontinuity field labels

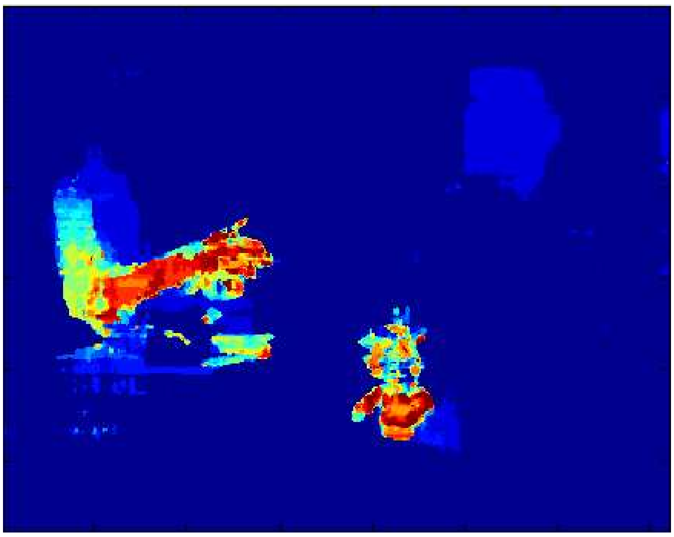

**Figure 6.** Motion field for 75-th frame of the Interview sequence

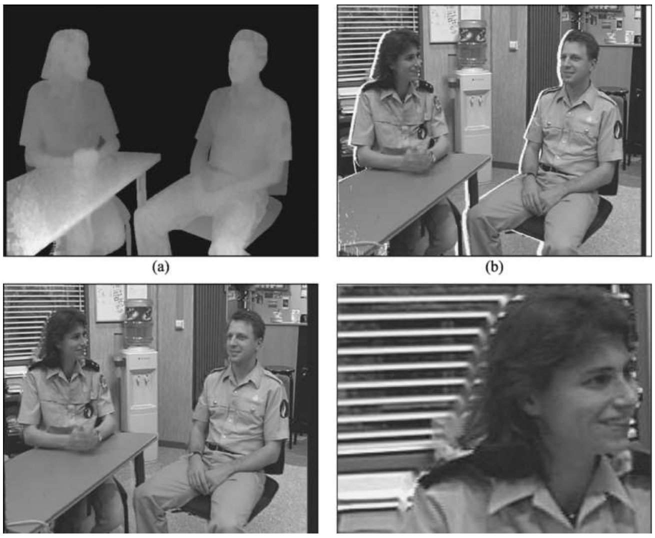

**Figure 7.** a.) Depth map of the Interview sequence b.) projected left-eye view c.) left-eye view after hole filling step d.) visible artefacts after hole filling

## 4. ARTEFACTS RELATED TO VIRTUAL VIEW GENERATION AND THEIR CORRECTION

Because of the difference in views, after projecting original central images using depth map, some surfaces, which are not visible from original view become uncovered in the virtual left and right-eye images. In the computer graphics literature, these uncovered areas are termed disocclusions. These areas have no associated textures after 3D warping because there are no available textures, neither in the depth information nor in the central image. Usual and the most simple technique to supress this kind of artefacts is so-called hole filling, where the newly exposed areas are obtained by averaging image values from neighboring locations.

Algorithms are evaluated using Interview sequence, first classical hole filling, than results obtained by using depth map smoothing and finaly our approach. In Fig. 8 shows central view, projected left-eye view, projected image after hole filling, and enlarged details of Interview sequence after filling.

White areas in Fig. 8 are the newly exposed areas after projection. Exposed areas are usualy located along the boundaries of objects and the right limit of the image. This is a direct consequence of the depth changes near object boundaries. Hole filling step, removes these artefacts efficiently, but significant artefacts still can be observed near the objects boundaries.

These artefacts can be significantly reduced by using assymetrical smoothing of the depth maps [5]. Here authors use Gaussian filtering with $\sigma_v = 10$ and $\sigma_h = 90$ of the depth maps to ameliorate the problem of areas in depth maps which create occlusions. After depth map smoothing large number of regions in the depth maps which are causing occlusions dissapear. Figure 8 shows the depth image after smoothing, virtual view created using smoothed depth map, and details of the virtual view. It can be observed that the number of the occluded areas is significantly reduced, but some artefacts in the geometry of the objects are also introduced. For example, the leg of the table is no longer straight.

First step in the proposed method is discrete wavelet transform. In this paper, we were using non-decimated wavelet transform, with the Daubechies db2 wavelet on four decomposition levels. In the second step all wavelet subbands are projected using the depth map. After this step, hidden areas appear in all subbands with non-defined values. Due to a sparsity of wavelet coefficients, missing coefficients can be interpolated much more efficiently, since the most of the wavelet coefficients are close to zero and the important details of the images are contained in the small number of significant ones.

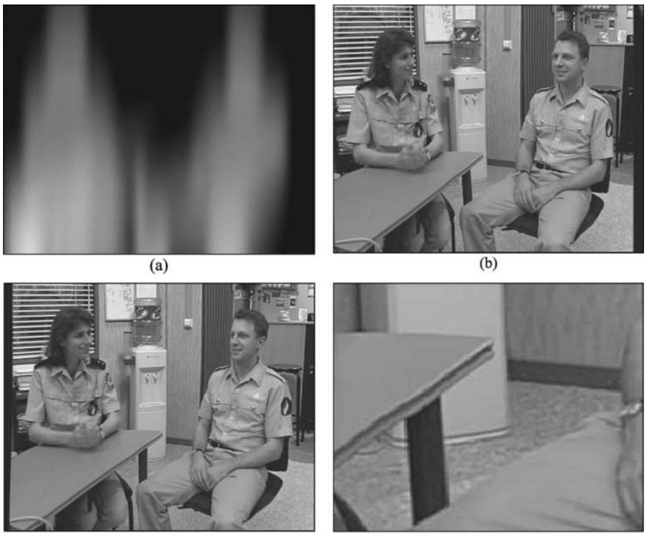

**Figure 8.** a.) Depth map of the Interview sequence b.) projected left-eye view c.) left-eye view after depth map smoothing d.) detail of the rendered view after depth map smoothing

Interpolation scheme for horizontal HL subbands first checks whether left and right neighbours of the occluded pixel exist, and then interpolates the coefficient using these two values. For vertical (LH) subbands upper and lower neighbours are checked and the non-existing coefficient is estimated by the average of neighbouring coefficients. If these do not exist, diagonal neighbours are checked and interpolated. This is done in an iterative manner, so the newly calculated coefficients can be used to correct other non-existing coefficients. Correction is done in three iterations. For the LL scale slightly different scheme is applied, since it contains low-pass filtered version of the original image and do not show regularitites as HL and LH band do. Here we check how precise existing neighbouring LL coefficients can be interpolated from neighbouring coefficients, and use pair of corresponding coefficients to interpolate the missing one. The main advantage of this method is that it does not introduce geometrical distortions as in the case of the depth map smoothing.

In the case when reliable motion estimates exist, estimated vectors are used for interpolation of the missing parts of the projected images. Both information from previous and current frame are used to find missing data. Fig. 9 illustrates the results of the proposed approach.

As can be seen in the Fig. 9, the proposed approach successfully removes occluded areas in the image while keeping the geometry and the depth map intact. This can be visible especially around the areas near edges of the objects. Since depth maps are not processed, depth resolution in the rendered stereo views remains the same as in original, as opposed to the reference approach of [5].

## 5. CONCLUSIONS

Main novelties in our algorithm are the use of wavelet coefficients for supression of occluded parts in the virtual view rendering and the use of motion vectors to further reduce the artifacts introduced by interpolation. One of the advantages of the presented approach is a novel, highly reliable motion estimation scheme, where the motion vectors estimated on the depth sequence are used to improve motion estimation on the luminance sequence and vice versa. We tested our method on multiple sequences which are typically used for comparison in the literature. The results demonstrate improvements over the best reported algorithms both in terms of hiding various artefacts and in terms of interpolating disoccluded areas.

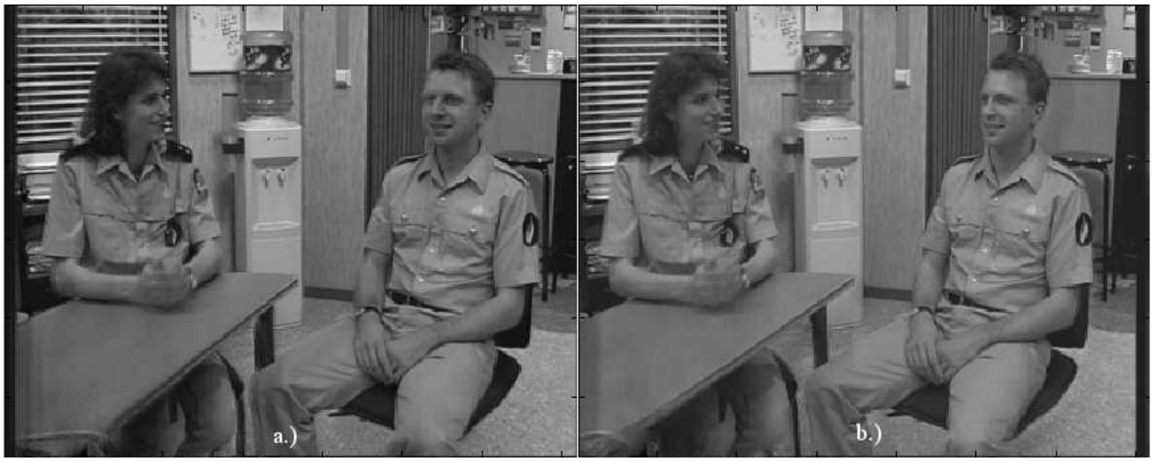

**Figure 9.** Rendered left a.) and right b.) eye view using the proposed method

## REFERENCES

1. I. Yuyama and M. Okui, *Three-Dimensional Television, Video, and Display Technologies*, Springer Press, Berlin, 2002.
2. Y. Luo, Z. Zhang, and P. An, "Stereo video coding based on frame estimation and interpolation," *IEEE Trans. on Broadcast.* **49**, pp. 14–21, 2003.
3. C. Fehn, "A 3d-tv approach using depth-image-based rendering (dibr)," *Proc. VIIP 03*, (Benalmadena, Spain), 2003.
4. J. Shade, S. Gortler, L. He, and R. Szeliski, "Layered depth image," *Proc. SIGGRAPH'98*, pp. 231–242, 1998.
5. L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3d tv," *IEEE Trans. on Broadcast.* **51**, pp. 191–199, 2005.
6. G. Alain, W. J. Tam, and L. Zhang, *Improving Stereoscopic Image Quality of Pictures Generated From Depth Maps*, Communications Research Centre Canada, Internal CRC report, Ottawa, Apr. 2003.
7. W. J. Tam, G. Alain, L. Zhang, T. Martin, and R. Renaud, "Smoothing depth maps for improved stereoscopic image quality," in *Conf. Three-Dimensional TV, Video, and Display III*, *Proc. SPIE* **5599**, pp. 162–172, (Philadelphia, U.S.A.), 2004.
8. S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, New York, 1995.
9. H. P. H. S. Kim, "Motion estimation using low-band-shift method for wavelet-basedmoving-picture coding," *IEEE Trans. on Image Processing* **9**(4), pp. 577–587, 2000.
10. V. Zlokolica, A. Pižurica, and W. Philips, "Wavelet-domain video denoising based on reliability measures," *IEEE Trans. on Circuits and Systems for Video Technology* **7**(3), pp. 477–488, 2006.