Abstract—In free viewpoint video (FVV) framework, a large number of viewpoints from limited number of views are generated to reduce the amount of transmitting, receiving and processing video data significantly. To generate a virtual view, the disparity among adjacent views or temporal correlation between different frames of the intermediate views are normally exploited. Those techniques may concern poor rendering quality by missing pixel values (i.e. creating holes) due to the occluded region, rounding error and disparity discontinuity. To address these problems recent techniques use inpainting, however, they still suffer quality degradation due to lack of spatial correlation on the foreground-background boundary areas. The background updating techniques with Gaussian mixture based modelling (GMM) can improve quality in some occluded areas, however, due to the dependencies on warping of background image and spatial correlation they still suffer quality degradation. In this paper, we propose a view synthesized prediction using Gaussian model (VSPGM) technique using the number of GMM models rather than the background image to identify background/foreground pixels. The missing pixels of background and foreground are recovered from the background pixel and the weighted average of warped and foreground model pixels respectively. The experimental results show that the proposed approach provides 0.50–2.14dB PSNR improved synthesized view compared with the state-of-the-art methods. To verify the effectiveness of the proposed synthesized view, we use it as a reference frame with immediate previous frame of current view in the motion estimation for multi-view video coding (MVC). The experimental results confirm that the proposed technique is able to improve PSNR by 0.17 to 1.00 dB compared to the conventional three reference frames.

Keywords—Depth image based rendering; spatial correlation; temporal correlation; interpolation; multi-view video coding; multi-view video plus depth

I. INTRODUCTION

Latest three dimensional (3D) technologies such as free viewpoint television (FTV) videos have become increasing interest for generating a more realistic impression of a scene in everyday applications such as teleconference systems. A large number of views with smaller baseline distance are required to facilitate this luxury, which increases transmission bandwidth and processing time for video data significantly. Therefore, the best policy would be to encode a subset of views and the receiver will synthesis the desired view from the limited decoded views. The most popular view synthesis technique is depth image based rendering (DIBR) which relies on depth map or the geometry of a scene [1]-[4].

DIBR technique takes a depth map together with texture image to synthesis a virtual view [3]-[6]. However, if the rendering views contain holes due to occlusion, it is not possible to overcome by using a single reference camera view [7][8]. Two adjacent cameras can cover relatively wider viewing angle. Therefore, adjacent views and their corresponding depth maps are used to minimize the occlusion problem for generating a virtual view [8]-[10]. However, in the practical application, a small numbers of views are transmitted due to the limitation of bandwidth. Therefore, the rendered view from a limited number of textures and depth maps would miss some pixels information [12]. The existing techniques based on a small number of views suffer from several problems such as occluded region, low precision rounding error and disparity discontinuity [11][13]. Most of the cases, an inpainting technique is used to recover occluded areas by exploiting spatial correlation [14]. The inpainting technique computes the priority of every pixel on the hole’s boundary and then copies the corresponding portion of the source patch in the target region of the target patch [14][15]. However, the success of the inpainting depends on how efficiently the priority of the hole is selected and on how the source patches selected for a target patch [14]. Moreover, the mixing up of the foreground and background priority patches can deteriorate the inpainting process. There are a number of techniques for overcoming those problems, however, due to the low spatial correlation in foreground/background boundaries this technique are not effective [16].

The above problem can be solved by treating foreground and background areas separately based on Gaussian mixture modelling (GMM) [17][18]. The techniques exploit temporal correlation using GMM-based background updating to recover missing pixels of the occluded areas. GMM was used to generate a background texture frame and then refine it by removing rotational foreground areas using clustering techniques with the corresponding depth image. The refine background image is used for recovering missing occluded static background pixels of the synthesised view. They used a conventional inpainting method for the missing pixels of static foreground areas. The experimental results show that the
background updating method provides a significant improvement compared to other methods including popular inpainting methods [18]. Due to the dependency on the warping of a background image and inpainting methods, if there is any error in background frame it propagates to the clustering technique as well. As they need hole-filling in two consecutive processes, any inaccuracy of the first step also deteriorates the quality of the final step. Moreover, the background pixel generation based on the mean value of the model sorted by the ratio of weight and standard deviation does not represent the recent changes of the pixel. As a result poor background pixel has been generated for their approach [19].

In the proposed VSPGM technique, we utilized the inherent capacity of the GMM to identify foreground and background pixels to recover occluded areas. To address the above mentioned problems, we use the number of Gaussian mathematical models representing a pixel rather than the end product i.e. background to identify background and foreground the pixel. According to the GMM technique only one model is introduced for a pixel if the pixel experiences similar intensities over the time, which indicates that the pixel is a background pixel. On the other hand, a pixel may experience background and foreground objects if a pixel is represented by more than one GMM models. Thus, our assumption is that the number of GMM models would be a good indicator to identify background/foreground pixels. In the proposed technique, we recover the missing pixels of the foreground and background areas from the adaptive weighted average of warped image and the background image and background image respectively. We apply GMM on the images of the interpolated view rather than on the adjacent view assuming that we have already synthesized previous images of the interpolated view. Thus, in the proposed method we can get better pixel correspondences. The experimental results confirm that the proposed technique could provide promising results.

View synthesized techniques recognized as a promising tool for rendering new views from multi-view video plus depth (MVD) for supporting advanced 3D video coding [20]. This technique provides an extra reference for multi-view video coding (MVC) by exploiting disparity among adjacent views. Due to the high similarity of the proposed synthesized view with the current view, these techniques provide better prediction compare to the conventional three references (i.e. two frames from adjacent views and the previous frame of the current view) system. To verify the effectiveness of the proposed synthesized view, we use two reference frames using the proposed synthesized view and the previous frame. The results show that we can improve the PSNR compared to the three reference scheme in multi-view compression.

The main contribution of this paper is to introduce a new hole-filling method based on the mathematical model using GMM instead of inpainting technique in the occluded areas of the interpolated views. The rest of this paper is organized as follows: section II describes the proposed view synthesis technique with details of foreground and background pixel modelling. Section III describes experimental results, while Section IV concludes the paper.

II. PROPOSED VIEW SYNTHESIZED PREDICTION TECHNIQUE

In this paper, we propose a VSPGM technique for improving the quality of virtual view for 3D video and FVV by using standard multi-view video sequences. In this technique, we have taken two texture images and corresponding depths maps and camera parameters as inputs. Then, we render a virtual view, but this view contains many cracks, ghosts and holes. To reduce these missing pixels of the occluded region, we model each pixel using the GMM with available previous frames of the interpolated view. We assume that when we interpolate n\textsuperscript{th} frame of a virtual view, 1 to (n-1)\textsuperscript{th} frames of the virtual view are available for the GMM. We classify each pixel as a foreground or background based on the number of models in the GMM. Pixel intensities of the occluded areas are filled based on either completely from the pixel of the background model or from the weighted intensity between the rendered image and the foreground model(s). The following sub-section describes the technique of interpolating virtual view(s) with GMM based hole-filling technique.

A. Interpolating Virtual View

Let $\Gamma_1$ and $\Gamma_2$ be the view 1 and view 2 texture images and $\Omega_1$ and $\Omega_2$ are the correspondence depth maps of the same scene captured by two cameras at the same time. Generally, depth maps represent distance of objects from the camera which are quantized into 256 different values where 0 and 255 represent the farthest and nearest distance respectively. The true depth values $Z$ are converted from the encoded depth map $\Omega$, the farthest depth in the scene $Z_{\text{max}}$, and the nearest depth in the scene $Z_{\text{min}}$ as by using (1) [10][12][21]:

$$Z = \frac{\Omega - Z_{\text{min}}}{Z_{\text{max}} - Z_{\text{min}}}$$

(1)

Then the disparity ($\delta$) between the reference view and the virtual view is calculated by using (2)

$$\delta = \frac{f \rho}{Z}$$

(2)

where $f$ is the camera focal length and $\rho$ is the baseline distance, i.e. the horizontal distances between the reference view and the virtual view position. After adjusting the true depth values and calculated disparity, the texture image is aligned in the new position [21][22]. However, this aligned texture contains many holes due to the quantization error, discontinuity of disparity and occlusion problem. To minimize the hole problems, we warped adjacent texture by using corresponding depth images to a virtual position [14][22][23]. After that, blending both warped images based on four conditions are as follows:
\[ \phi = \begin{cases} 
(1 - \beta) \Gamma_1' + \beta \Gamma_2', & \text{if no holes in } \Gamma_1' \text{ and } \Gamma_3' \\
\Gamma_1', & \text{if no holes in } \Gamma_1', \text{ but holes in } \Gamma_3' \\
\Gamma_3', & \text{if no holes in } \Gamma_3', \text{ but holes in } \Gamma_1' \\
0, & \text{if holes in } \Gamma_1' \text{ and } \Gamma_3'. 
\end{cases} \]  

(2)

where \( \Gamma_1' \) and \( \Gamma_3' \) are the view 1 and view 3 warped images, \( \phi \) is the rendered view and \( \beta \) is a weighted factor (to generate a middle view, the value of \( \beta \) is 0.5). This procedure reduces the number of holes, but does not help to recover all missing pixel intensities. To recover missing pixel, we model each pixel by using GMM technique.

**B. GMM Technique**

Each pixel position of a scene is modeled independently by a mixture of \( K \) (generally 3 models [24][25]) Gaussian distributions. Let at time \( t \) a pixel intensity of \( k \)-th Gaussian representing, \( \mu_{k,t} \), with mean, \( \mu_{k,j} \), weight, \( \omega_{k,t} \), and the variance, \( \sigma^2_{k,j} \) such that \( \sum \omega_{k,j} = 1 \). The fixed initial parameters such as standard deviation, \( \sigma_k = 2.5 \), initial weight, \( \omega_{k,t} = 0.001 \) and learning rate, \( \alpha = 0.1 \) are used in the proposed technique. For balancing the contribution between present and previous values of variance, mean, and weight, a learning parameter \( 0 < \alpha < 1 \) is used [19][26].

After initialization, at the time \( t \) for every new observation, the new pixel intensity is \( X_t \) such that

\[ |X_t - \mu_{k,j}| \leq 2.5\sigma_{k,j} \]

for the first match against existing models. If a model matches, associated parameters are updated as follows:

\[ \mu_{k,j} \leftarrow (1 - \alpha) \mu_{k,j-1} + \alpha X_t; \]  

(3)

\[ \sigma^2_{k,j} \leftarrow (1 - \alpha) \sigma^2_{k,j-1} + \alpha (X_t - \mu_{k,j})^T (X_t - \mu_{k,j}); \]  

(4)

\[ \omega_{k,j} \leftarrow (1 - \alpha) \omega_{k,j-1} + \alpha, \]  

(5)

and the weights of the remaining Gaussians are updated as

\[ \omega_{k,j} \leftarrow (1 - \alpha) \omega_{k,j-1}. \]  

(6)

After that the weights are re-normalized among all models so that the total value is 1. On the other hand, if a model doesn’t match, a new model is introduced using initial parameter values. If it has already reached at the maximum allowable number of models, the new model replaces an existing model based on the value of \( \omega/\sigma \). When a pixel intensity of a color \( (c) \) satisfies a model \( (k) \), we also keep the pixel intensity as the recent value \( \Delta_{k,j}^c \) of the corresponding model and color. The value of \( \Delta_{k,j}^c \) will be used in the missing pixel recovering technique.

Thus, each pixel can be represented by a number of Gaussian models. If a pixel represents a static background over the time, then it might have only one model, on the other hand, if a pixel experiences foreground/background, it might have more than one model where one model represents a stable background and other models might represent foreground/background. The model with the highest value of \( \omega/\sigma \) representing the background and other models represent foreground. As the GMM can successfully capture foreground and background pixel intensities by exploiting temporal correlation, missing pixels of an occluded area are successfully recovered.

**C. Hole Filling Approach**

In the GMM technique, the Gaussian models of a pixel are always ordered based on the \( \omega/\sigma \) in descending order, assuming that the top Gaussian will provide most stable background. In the proposed technique if a pixel is modelled using only one Gaussian model, the pixel intensity of the interpolated final image \( \Psi_t^c \) is taken from the recent value i.e. \( \Delta_{k,j}^c \) of the model. Otherwise, the pixel intensity of the interpolated image is taken as a weighted average from the rendering image \( \Phi_t^c \) (i.e. output image after warping) and the recent value of the model which provides lowest value in terms of \( \omega/\sigma \). The detail of the interpolated image recovering technique using GMM is described below [27]:

**Case 1:** If a pixel has only one model for a given color, we assign the recent value \( \Delta_{k,j}^c \) of the color of the interpolated image using

\[ \Psi_t^c = \Delta_{1,j}^c. \]  

(7)

**Case 2:** If a pixel has two models for a given color a pixel experiences foreground/background, therefore, we choose a weight factor \( (\lambda) \) for selecting the fraction of the rendering image and the recent value of the second models as follows:

\[ \Psi_t^c = \lambda \Phi_t^c + (1 - \lambda) \Delta_{2,j}^c. \]  

(8)

**Case 3:** If a pixel has three models for a given color, we have used equation (9) for selecting fraction of the interpolated image as follows:

\[ \Psi_t^c = \lambda \Phi_t^c + (1 - \lambda) \Delta_{3,j}^c. \]  

(9)

**III. VIEW SYNTHESIS FOR MVC**

Adjacent views of multi-view video sequences are captured by multiple cameras with slightly different angles. Therefore, there are disparities among different views. Moreover, co-located pixels/blocks at different instant of the same views are predicted by motion estimation technique. However, finding co-located pixels/blocks on different frame by using motion estimation and disparity estimation is time consuming [28]. Therefore, reduction of computation for searching motion parameters such as motion vector is an important aspect of
current research [29]. Thus, the best policy is reducing the number of reference views. Traditionally, three references such as already encoded frames of adjacent views (reference frames 1 and 2) and previous frame of the current view (reference frame 3) are used to encode each frame of dependent views [20]. In this technique, a disparity \( \delta \) is used to find a current block (CB) \((X_c, Y_c)\) on adjacent reference views \( (X_{r1}, Y_{r1}) \) and \( (X_{r2}, Y_{r2}) \) where \( X_{r1} = X_c + \delta \) and \( X_{r2} = X_c - \delta \). This method only considers the horizontal component as multi-view video sequences are rectified [9]. Furthermore, motion vectors are predicted to find a CB on the previous frame of the current view i.e. \((X_{r3}, Y_{r3})\) [9][13]. Instead of typical approaches, we will use a view synthesis technique to generate a synthesized current frame i.e. reference frame 2. This synthesized frame is almost similar in terms of object position and its motion to the expected current frame. If we would consider two references such as reference 1 and reference 2 as shown in Fig. 1, it provides better prediction compared to the traditional approaches.

**IV. EXPERIMENTAL RESULTS**

In this section the performance of the proposed VSPGM technique analysed and compared with standard techniques such as HTM Renderer (HEVC Test Model) [30], inpainting technique [14] and background update technique [18] based on peak-signal-to-noise-ratio (PSNR). We apply the inpainting method on the rendering image created after warping. We also use the same technique and then apply background update technique and the VSPGM technique for the refinement of the interpolated view. Fig. 2 shows the PSNR comparison between VSPGM, HTM, inpainting and background update techniques. The figure reveals that the proposed technique outperforms the existing hole-filling technique for all video sequences. The improvement range varies from 0.1dB to 9.20dB for HTM, 0.50dB to 7.66dB for inpainting and 0.50dB to 2.14dB for background update technique respectively.

Fig. 3 illustrates the frame differences between the original images and the reconstructed images of the proposed and the state-of-the-art methods. The figure reveals that the proposed technique is able to generate more similar images compared to the state-of-the-art methods. To understand the contribution between the GMM models and the rendering image in the proposed scheme to reconstruct the final interpolated image, we analysis PSNR against different values of \( \lambda \) in Fig. 4. The figure reveals that the contribution of rendering image and the GMM based pixel intensity in the foreground areas varies for different sequences. It can be easily observed from the figure that both rendering image and GMM have some contribution to generate an interpolated image for each video. Note that if we get the maximum PSNR value of a given image where the value of \( \lambda \) is 1.0, it means that the pixel intensities of the interpolated image for foreground is entirely taken from the rendering image. However, the background pixel of the interpolated image is always taken from the recent pixel value of the stable background model in the GMM. Newspaper video sequence has maximum foreground areas as well as occlusion areas which are not able visible in the warped images. Thus, increasing the contribution of the warped images reduce overall PSNR. The content of the Lovebird1 video sequence i.e. foreground and background areas are almost balanced, therefore the contribution of the warped images and GMM model are varied slightly with varying the values of \( \lambda \). Poznan Street video sequence has less occlusion areas, therefore it increases contribution of the warped images which has less occlusion areas and improved PSNR with increasing the values of \( \lambda \).

Fig. 5 illustrates the subjective quality for Newspaper video sequence. Fig. 5 (a) shows the original images, i.e. 11th original frame of the virtual view and the green rectangular boxes are used to mark the cropped and zoomed portion which are shown in Fig. 5 (b) and (c). Similarly, Fig. 5 (d), (g), (j) and (m) shows the interpolated view by HTM, inpainting, background update and VSPGM technique and Fig. 5 (e), (f), (h), (i), (k), (l), (n) and (o) shows corresponding cropped and zoomed images. The figures reveal that the proposed method is able to generate a better virtual view compared to the state-of-the-art methods.
Fig. 3: Comparison of frame differences between the original image (11\textsuperscript{th} frame) and the corresponding generated virtual image using standard multi-view video sequences by the proposed method and three state-of-the-art methods.
To understand the recover areas in the interpolate virtual image taken from the rendering image and the background image, we provide pixel categorization based on the source of refinement. In this regard Fig. 5 shows the pixels that are selected from the background image (black portion) and rendered image (non-black portion) for generating final virtual view for the Newspaper and Lovebird1 video sequences. The figure demonstrates that the moving areas i.e. foreground areas are normally taken from rendering image and the static background/occluded areas are taken from the background image. The results confirm our hypothesis on the effectiveness of the foreground/background pixel identification using the number of GMM models.

Table 1: PSNR comparison for the proposed MVC scheme

<table>
<thead>
<tr>
<th></th>
<th>Three references</th>
<th>Two references</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Newspaper</strong> (PSNR)</td>
<td>38.24</td>
<td>39.10</td>
</tr>
<tr>
<td><strong>Lovebird1</strong> (PSNR)</td>
<td>37.41</td>
<td>37.58</td>
</tr>
<tr>
<td><strong>Poznan Street</strong> (PSNR)</td>
<td>35.74</td>
<td>36.73</td>
</tr>
</tbody>
</table>

To encode different resolutions and wide range of video contents for different views each frame is divided into the number of blocks with various sizes such as 8x8, 16x16, 32x32 and 64x64 [9] pixels and the search length become 8, 16, 32, 64 and 128 pixels. In our experiment, we have considered 32x32 pixel block sizes and 64 pixels search lengths. Due to the better prediction of synthesized view, the proposed techniques provides better PSNR compare to the conventional approaches as shown in Table 1. It reveals that the average PSNR improvement for two references is 0.70dB.

V. CONCLUSION

In this paper, we present a new VSPGM technique that exploits temporal correlation for improving the quality of synthesised views compared to the existing methods. Virtual images are generated from a texture image and its corresponding depth map. Interpolated virtual images contain many holes due to the occlusion and rounding error problem. To address these issues, inpainting and/or background updating methods are used for the most cases for refining the virtual image. Due to the lack of spatial correlation in the background-foreground areas and warping/clustering problems of existing methods, they fail to provide good virtual views. In the proposed method, we recover the missing pixel intensities using the number of Gaussian mixture-based models which have the capacity to identify foreground and background pixels based on the temporal correlation. The background pixels are recovered from the stable background model and the foreground pixels are recovered as the weighted pixel intensities of the rendering image and the pixel intensities of the foreground models from GMM. The experimental result shows that the proposed method improves 5.2dB, 1.60dB, and 5.84dB PSNR on average compared to the inpainting, background update and HTM techniques respectively. To evaluate the performance of the proposed technique we used synthesised frame as a reference with immediate previous frame of the current view for MVC, it improves 0.70dB PSNR on average compare to the standard techniques.
Fig. 5: Original image (a), synthesised images (d, g, j and m), crop and zoom images (b, c, e, f, h, i, k, l, n and o) for *Newspaper* video sequence by the VSPGM method and three state-of-the-art methods.
REFERENCES