A NEW FAST MOTION ESTIMATION AND MODE DECISION ALGORITHM FOR H.264 DEPTH MAPS ENCODING IN FREE VIEWPOINT TV

G. Cernigliaro[a], M. Naccari[b], F. Jaureguizar[a], J. Cabrera[a], F. Pereira[b], N. García[a]
[a]Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, Madrid, Spain
[b]Instituto Superior Técnico, Instituto de Telecomunicações, Lisboa, Portugal

ABSTRACT

In this paper, we consider a scenario where 3D scenes are modeled through a View+Depth representation. This representation is to be used at the rendering side to generate synthetic views for free viewpoint video. The encoding of both type of data (view and depth) is carried out using two H.264/AVC encoders. In this scenario we address the reduction of the encoding complexity of depth data. Firstly, an analysis of the Mode Decision and Motion Estimation processes has been conducted for both view and depth sequences, in order to capture the correlation between them. Taking advantage of this correlation, we propose a fast mode decision and motion estimation algorithm for the depth encoding. Results show that the proposed algorithm reduces the computational burden with a negligible loss in terms of quality of the rendered synthetic views. Quality measurements have been conducted using the Video Quality Metric.

Index Terms— Free Viewpoint Video, H.264, Mode Decision, Motion Estimation, View+Depth, depth maps.

1. INTRODUCTION

Free Viewpoint Video (FVV) and 3DTV [1] are the next generation of video paradigms whose goal is the involvement of the observer thanks to the 3D perception of the scene. In order to display the 3D scene and to capture the real world a high amount of points of view is needed, therefore a certain number of cameras is used. The possibility of reducing the number of cameras is given by generating synthetic points of view through the use of techniques which are able to create a virtual sequence placed between two real videos. The view generation techniques [2] are interpolation methods which, apart from the two real video sequences, need to know the depth information of the scene. The sequences used in this scenario are called View+Depth sequences and every frame is represented by a traditional visual information frame with its corresponding depth map. A depth map is a representation of the scene in grey levels where every value represents a different distance with respect to the camera: light values are assigned to objects close to the camera and, on the contrary, dark values are used to represent the far objects. As in FVV this information is used to represent the 3D scene, the video coding scenario must be subject to some changes and must be adapted to admit this new type of sequences in order to obtain a good compression and to not waste important computational resources in unnecessary processes.

The depth additional information makes greater the amount of data to compress and, as consequence, also the computational burden of the encoding process. As the depth is a different representation of the same scene recorded by the view, there could exist some similarities between the motion characteristics of the view and of the depth. The goal of this study is to design a fast Motion Estimation/Mode Decision (ME/MD) algorithm, based on H.264/AVC [3], able to exploit this similarities and, as consequence, to reduce the complexity of the encoding process of the depth. In other studies, the investigation was focused on the use of the motion information (modes and Motion Vectors (MVs)) of the view, to approximate the estimation of the motion information of the depth. Hewage *et al* [4] designed a method where a loss of the depth motion information could be concealed by using the corresponding data from the view, copying it directly during the decoding process, demonstrating that it could be a good solution to estimate the depth motion characteristics. However this solution is oriented to a scenario where videos are encoded in a traditional way and only some loss has to be recovered. In the work of Oh and Ho [5], it has been analyzed more in detail whether the motion information of the view is useful to encode the depth. The study has been carried out by evaluating the differences between the MVs of the view and the depth. The obtained conclusions show that a good reliability is not always guaranteed because it depends on the characteristics of the sequences, consequently different strategies are needed during the MD and ME stages. Taking into account this background, to reach the goal of this work, a more in depth study is necessary to examine when strategies as the direct copy of the modes and MVs are reliable and when other approaches are needed. For this reason the first part of this work is oriented on the analysis of the motion characteristics of the depth and on the comparison of them with the motion characteristics of the view. Later on, this study will represent the basis on where the novel ME/MD algorithm will be built.

All the work presented in this paper is oriented to the FVV environment. In this scenario, depth sequences are never shown, but they are used to create synthetic views which are displayed and, depending on the compression algorithms applied on the depth, they could be represented with variable qualities. The quality of depth map compression is usually evaluated by using objective metrics, e.g. the well known Peak-Signal-to-Noise-Ratio (PSNR), applied directly on the depth map data. However, as stated above, depth maps are never showed to the viewer and therefore a direct assessment over them does not allow to fully assess the actual quality perceived by a human observer. Therefore, in this paper we proposed to determine the depth map quality by evaluating the quality of the synthesized view by mean of the Video Quality Metric (VQM) which shows a good correlation with the perceived video quality [6].

The paper is structured as follows. Section 2 describes the first part of the work in which the motion characteristics of the View+Depth sequences are analyzed to find the best strategy for the depth maps encoding process. In Section 3, the encoder structure and the proposed algorithm are explained. Section 4 shows the experimental results of the application of the algorithm, and finally, Section 5 presents the conclusions.

2. VIEW+DEPTH JOINT MOTION ANALYSIS

In order to reliably use the view motion information to encode the corresponding depth, it is necessary to carry out an analysis of the motion characteristics of the View+Depth sequences in a traditional scenario, encoding the view and the depth independently and using ME and MD algorithms implemented in the Reference Software (RS)[7].

This work has been partially supported by the Ministerio de Ciencia e Innovacion of the Spanish Government under project TEC2010-20412 (Enhanced 3DTV).

(a) View Frame

(b) Depth Frame

Fig. 1. Frame #3 of Beergarden with its corresponding depth map.

2.1. View motion field analysis

The first analysis has been carried out in order to know whether the ME evaluated for the view is a good approximation to encode the corresponding depth. Therefore it is necessary to understand if, when using the MVs of the view on the depth, depth motion is well approximated. During the ME stage, in H.264, the best reference area to predict and encode a Macro-Block (MB), or its subsections, is searched. The resulting MV is the difference, in pixels, between the position of the MB and its reference. If the motion information evaluated for the view is a good approximation to encode the depth, by evaluating a Motion Compensation (MC) of it through the MVs of the view, the prediction should result acceptable and the differences between the MBs and the areas so referenced should be reduced. In order to provide an intuition about the conclusions carried out through this study, it has been used the sequence Beergarden because of its emphasized borderlines between areas with different characteristics. The analysis presented in this section has been made on the frame number 3 of the 5th view of the Beergarden sequence (Figure 1).

The view has been encoded by using only the Inter prediction in order to obtain MVs for every MB. It has been used the Mean Square Error (MSE) metric between the original depth frame and the reconstruction of the reference depth frame compensated with the MVs of the view frame. Figure 2 represents the MB-wise MSE computation. We can observe that almost all the MBs have an MSE close to zero except from the areas where the edges of the silhouettes are present: in these regions the prediction is not so accurate. So the motion of the view can be used to predict the motion of the depth, but, where an edge of an object is present, a different strategy should improve the performance.

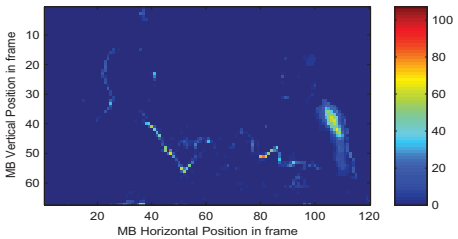


Fig. 2. MB-wise MSE between the depth frame and the previous depth frame compensated with the MVs of the view frame.

2.2. View+Depth modes analysis

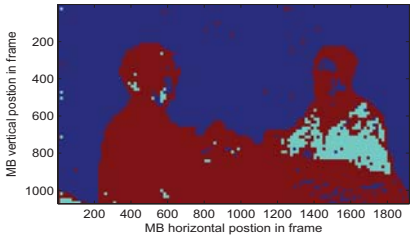
The good approximation of a depth area through the MVs of the same region of the view does not imply that its modes should also be the best choice to encode it. Therefore the second part of this View+Depth joint motion analysis focuses on the difference of the modes of a view and its associated depth.

2.2.1. View Modes analysis

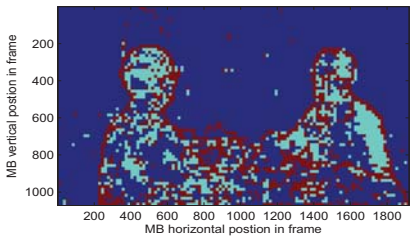
In this case, the encoder configuration has been set by also enabling the Intra modes. This is due to the texture differences between depth and view which could need different prediction strategies. Figure 3 shows a graphic representation of the modes evaluated by a traditional H.264 encoder for every MB of the 3rd frame of the view number 5 and of its depth. Figure 3.a shows the best modes to encode the view frame: the background areas are almost all encoded by using the Skip mode whereas the foreground areas are almost all encoded by using the other Inter modes. Only a reduced area of the foreground is encoded as Intra due to the uniformity of the texture which causes that a spatial-based prediction can provide better compression efficiency than a temporal-based one.

2.2.2. Depth Modes analysis

Applying the same analysis on the corresponding depth, the result shows different characteristics. The selected modes of the foreground area of the depth are different with respect to those in the view. Figure 3.b shows how this region of the depth is encoded by using a higher number of Intra MBs than the previous case (view). The reason that why there is a difference in the mode selection it is that the depth is a representation of the position of the objects in the 3D scene, so a plane surface, parallel to the camera, will be uniform in the depth domain regardless of the same region in the view. This situation, in the depth, could need different prediction strategies than for the view and, as the results show, an Intra prediction obtains best performances. The conclusion carried out by this part of the analysis is that the mode evaluated for a view MB is not always reliable to encode the corresponding depth MB and it is not possible to reuse this information in all the situations.



(a) View Frame



(b) Depth Frame

(c) Legend

Fig. 3. MD evaluated by H.264 for the frame #3 of the 5th view of the sequence Beergarden and of the corresponding depth map.

3. DEPTH FAST MOTION ESTIMATION AND MODE DECISION ALGORITHM

According to the analysis made in Section 2 the proposed fast ME and MD algorithm takes into account the similarities between the motion of a view and of its depth to make faster the ME/MD stages of the depth encoding process, but it also considers the differences which make that the reuse of the view information is not always reliable. This Section describes the steps of the method and the different strategies applied depending on the cases. The ME and MD stages, in H.264, are applied independently on every MB, so the application of the proposed algorithm will be described for one MB. Figure 4 shows all the steps of the fast ME/MD method.

The accomplished analysis came to the conclusion that the presence of an object edge could represent a critical situation in considering reliable the motion information of the view in the encoding of the depth. *Step 1*, which is the first stage of the proposed algorithm, has the task to find these areas using a Sobel edge detector. When an edge is detected the reuse of the motion information evaluated for the view can not give good performances, therefore the use of traditional, H.264 based, ME/MD algorithm is more desirable. Whether an edge is not present it is possible to go to the *Step 2*. At this stage the analysis of the view motion information starts. *Step 2* begins by reading the mode of the MB of the view in order to decide if it is possible to use it to predict the mode of depth MB. If the view MB has been encoded as Intra, the prediction has been evaluated according to the texture characteristics of the view frame. Section 2.2 has explained that the texture of a view and of its depth, although they belong to the same object, could have different characteristics which need different encoding approaches, so, also in this case, it is not possible to rely on the reuse of the view motion field and a traditional ME/MD algorithm is considered. When the view MB is an Inter MB the following step is reached. *Step 3* only checks if the view MB has been encoded as Skip or through another Inter mode. If the Skip mode has been detected there are not MVs to consider and the algorithm goes directly to the *Step 5*, otherwise the *Step 4* needs to be processed. At this stage the algorithm reads the MVs of the view MB and evaluates the ME only according to the view mode and the view MVs. *Step 5* has the task to test the Skip mode of the depth MB and *Step 6* tests the Inter modes. The reason for which Skip and Intra modes are always between the candidates is due to the consideration achieved through the analysis made in Section 2 where it has been shown that there is not always a relationship between the mode of the view MB and the mode of the corresponding depth MB. It has been demonstrated that the different use of these modes is due to the different texture characteristics between the two types of video (view and depth) so, as it is not predictable the solution with best performance, it is preferable to always consider these modes between the candidates. Summarizing, *Step 1* and *Step 2* are able to detect when the motion information of the view is reliable and, in case of not reliability, apply a traditional ME/MD algorithm. *Step 3* checks if the view MB has been encoded as Skip or not, to organize the different candidates. *Step 4* evaluates the Inter mode and the Inter MVs of the view MB if this is Inter not Skip; this step is not considered when the view MB is Skip. The last two steps (*Step 5* and *Step 6*) are always processed and are in charge of testing Skip and Intra modes.

4. EXPERIMENTAL RESULTS

As it has been explained in the introduction (Section 1), depth is never displayed but it is used to generate the synthetic views. Therefore, the best way to evaluate the performance of a coding algorithm designed for depth sequences is to compare the synthetic views generated through the depth encoded by using the proposed method and by a traditional H.264 encoder. VQM metric has been used to assess the quality of the synthesized view given its good correlation with the perceived video quality by a human observer [6].

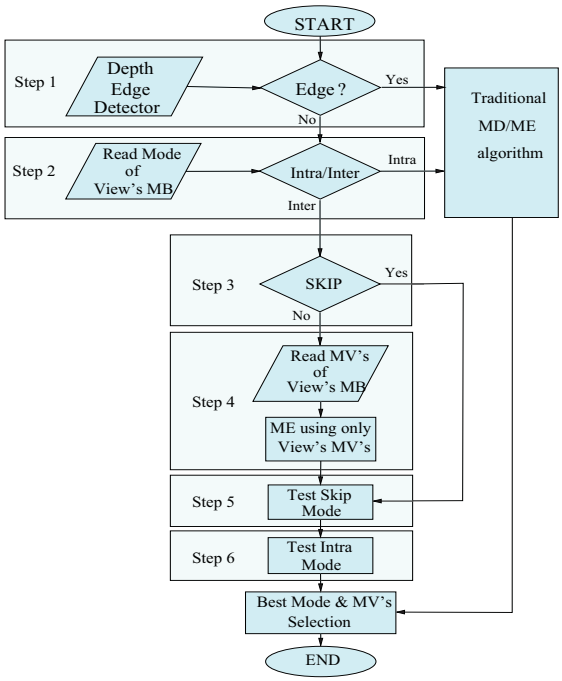


Fig. 4. Flowchart of the proposed algorithm.

4.1. Settings of the experiments

To encode the view sequences the used H.264 profile it has been the Main Profile of the Reference Software (RS). The encoder has been configured with an IPPP... Group of Pictures (GOP) with an Intra refresh every second. The enabled modes in the coding process have been all the possible Inter and Intra modes, the used MD algorithm was the High Complexity method and the ME algorithm was the Full Search made on windows of 32 pixels (default setting of the RS). In order to have a correct comparison, the configuration of the encoders used to compress the depth maps, in a traditional H.264 fashion, was the same. The software used was the RS version 17.1 [7]. The results have been obtained by processing the first 100 frames of the views number 5 and 6 of the sequence Beergarden (frame-rate = 25 fps) and the views number 2 and 4 of the sequence Newspaper (frame-rate = 30 fps).

VQM is designed to measure the perceptual differences in qualities between an original and a processed video. The perceptual differences are represented by values in the range [0,1], where zero means excellent quality compared to the original. In this experiments, the values have been obtained by comparing the synthetic views, generated through the depths decoded after the compression, with the synthetic views obtained with the original depths. All the synthetics views have been generated with the View Synthesis Reference Software (VSRS)[2].

As explained in Section 3, the first step of the proposed algorithm is an edge detection. A Sobel operator, implemented using OpenCV [8], has been used to perform this stage. The presence of an edge causes the not reliability of applying the motion information of the view on the encoding process of the depth and, as in this case a traditional H.264 method is used, the charge of the ME and of the MD stages will be greater. To make the process lighter, it is necessary not to consider all the edges, so, after the use of the So-

bel operator, the application of a threshold in order to eliminate the not significant edges is convenient. The employed thresholds have been decided through an empiric observation of the performances for every sequence.

4.2. Results Discussion

The goal of the algorithm proposed in this paper is the reduction of the computational burden of the ME and MD stages in the encoding process of the depth maps. Therefore, in order to evaluate the reduction complexity, the experimental analysis is also oriented to the evaluation of the differences in terms of coding time between the proposed fast method and a traditional H.264 encoder. Figure 5 displays the comparison of the quality performance for the sequence Beergarden. The considered rate is the sum of the rates used to encode the two depth sequences of the two views needed to generate the synthetic view. The result shows that the proposed method reaches the traditional H.264. The coding time comparison, represented in Table 1, demonstrates that this good performance is obtained with a remarkable difference in terms of time which is due to a reduced complexity. Figure 6 and Table 2 show that for other type of sequences (Newspaper) the quality is comparable to an H.264 encoder and the saved time is also considerable. The differences in the saved time between the two cases are due to the different thresholds used in the edge detection stage. In Beergarden a higher threshold has been used, so a fewer number of edges is considered and a traditional H.264 based ME/MD approach is needed in less MBs. As a consequence, the computational complexity is lower. As it has been explained in the previous section, the thresholds used have been obtained after an empiric analysis of the performances. The future goal of the study presented in this paper it is the achievement of the best performance with a threshold evaluated though an automatic method according to the characteristics of the sequences.

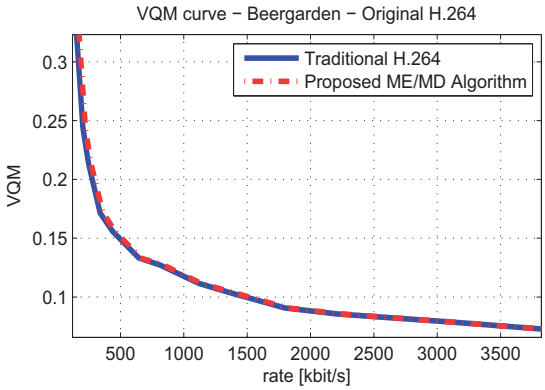


Fig. 5. Comparison of the performance of the proposed method vs a traditional H.264 encoder (sequence Beergarden).

5. CONCLUSIONS

In this paper a new fast algorithm to perform the ME and the MD stage for the encoding process of depth sequences in a FVV environment has been proposed. The motion characteristics of the View+Depth sequences have been analyzed in order to prevent critical situations and to optimize the strategies to obtain the best performances. Considering the particular characteristics of the considered scenario, which is quite different from a traditional video coding scenario, the VQM metric has been introduced as quality metric in order to analyze the performance of the algorithm in the best conditions. The results obtained by the proposed algorithm show that it is possible to reach optimum levels of quality saving a considerable quantity of computational charge and, consequently, coding time.

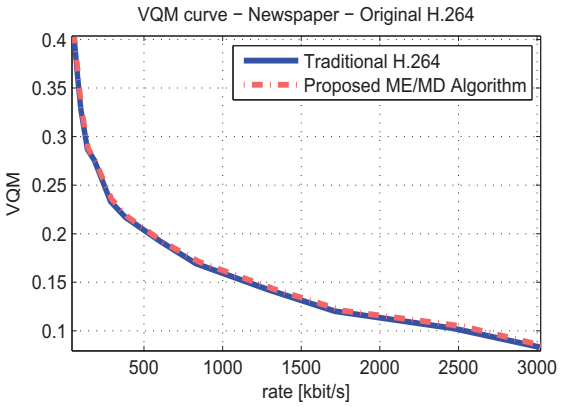


Fig. 6. Comparison of the performance of the proposed method vs a traditional H.264 encoder (sequence Newspaper).

Beergarden

QP	20	22	25	27	30	32
Saved time (%)	62.6	62.5	61.5	60.4	59.3	58.9
QP	35	37	40	42	45	47
Saved time (%)	57.2	55.6	56.6	56.1	56.8	55.9

Table 1. Saved encoding time by the proposed method with respect to the H.264 standard for the sequence Beergarden.

Newspaper

QP	20	22	25	27	30	32
Saved time (%)	29.2	28.9	29.2	30.4	29.1	29.7
QP	35	37	40	42	45	47
Saved time (%)	30.6	30.6	29.9	30.5	30	29.8

Table 2. Saved encoding time by the proposed method with respect to the H.264 standard for the sequence Newspaper.

6. REFERENCES

[1] Aljoscha Smolic, "An overview of 3d video and free viewpoint video," *Lecture Notes in Computer Science*, vol. 5702, pp. 1–8, 2009.

[2] Patrick Lopez Dong Tian, Po-Lin Lai and Cristina Gomila, "View synthesis techniques for 3d video," in *Advances in Image and Video Technology*. SProc. SPIE, 2009, vol. 7443.

[3] Iain E G Richardson, *H.264 and MPEG-4 Video Compression*, John Wiley & Sons, 2003.

[4] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, and A.M. Kondoz, "A novel frame concealment method for depth maps using corresponding colour motion vectors," in *3DTV Conference*, May 2008, pp. 149 –152.

[5] Han Oh and Yo-Sung Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," in *Advances in Image and Video Technology*. Springer Berlin / Heidelberg, 2006, vol. 4319 of *Lecture Notes in Computer Science*, pp. 898–907.

[6] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, September 2004.

[7] Fraunhofer-Institute HHI, "H.264/avc jm reference software (jm 17.1)," 2010.

[8] "Opencv (open source computer vision)," 2010.