An objective quality metric for panoramic videos

Sandra Nabil, Frédéric Devernay, James L. Crowley

An objective quality metric for panoramic videos

Sandra Nabil[1], Frederic Devernay[1], and James Crowley[1]

[1]INRIA Grenoble Alpes

***Abstract*—The creation of high quality panoramic videos for immersive VR content is commonly done using a rig with multiple cameras covering the required scene. Unfortunately, this setup introduces both spatial and temporal artifacts due to the difference in optical centers as well as the imperfect synchronization between them. Therefore, designing quality metrics to assess those videos is becoming increasingly important. Using traditional image quality metrics is not directly applicable due to the lack of a reference image. In addition, such metrics do not capture the geometric nature of such deformations. In this paper, we present a quality metric for panoramic video frames which works by computing pair-wise quality maps prior to blending and fusing them to obtain a global map of potential errors. Our metric is based on an existing one designed for novel view synthesized image, which is a similar problem to image stitching. Results show that applying this quality metric offers a practical way to assess panoramic video frames which do not have a reference. It also consolidates the similarity of the artifacts produced from novel view synthesis algorithms and those produced in the process of image and video stitching.**

I. Introduction

Panoramic videos have become an important tool for providing immersion in virtual reality (VR) environments. The richness of 360° panoramic videos makes it possible to capture scenes where the user can experience presence by looking through a head-mounted display. However, this realism can immediately be broken if the user spots any visual artifact.

One of the most disturbing visual error phenomena falls into the category of parallax errors which appear in the form of discontinuities, deformations or ghosting, as seen in figure I. This kind of distortion is nearly unavoidable in panoramic videos which are usually created using a panoramic rig with multiple cameras each covering a large field of view with a certain overlap with one or more other cameras. Traditional image quality metrics are not well suited for this problem due to the lack of a reference image. In addition, such metrics do not capture the geometric nature of such deformations. Establishing a usable quality measure to quantify these defects is fundamental for evaluating and comparing various stitching algorithms.

To overcome these limitations, we suggest to use a quality metric originally designed for novel view synthesized images [1]. The similarity between both domains, which fall into the category of IBR (Image-Based Rendering), makes them suffer similar artifacts. We propose a solution to the lack of a reference image by calculating the metric between overlapping regions of pair-wise matches in the original views prior to blending. We then construct a global map for the whole panorama to which we apply a weighting mask that accentuates the errors appearing around the boundary of two given views. Results show that this approach successfully spots potential errors giving more importance to the ones that are more visible according to the human visual system (HVS).

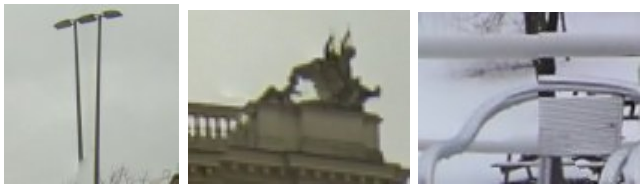

Fig. 1. Examples of parallax errors in panoramic videos.

The rest of the paper is organized as follows: section 2 provides is overview of related work, the new suggested approach in section 3 and finally the results in section 3. We end the discussion by the conclusion and future work in the last section.

II. Related Work

In the previous section, we talked briefly about why traditional quality metrics are not well suited for panoramic image quality. The fact that there are multiple source images to a single panoramic output with no ground truth to compare makes it difficult to establish a full reference quality metric. In addition, the processing of the input images includes various geometric transformations causing not only photo-metric distortions but more importantly structural ones.

Little work has been been published on panoramic images assessment. Recently, Yang et al [2] describe a method which assesses panoramic image quality based on the variation of flow field between two given scenes weighted by a salience map in addition to a structure histogram. The method seems effective, however it depends on the per-pixel motion field which might be erroneous itself. It also depends on a certain camera setup to obtain a ground truth image. Others have focused on the user experience such as in [3] who provided a subjective quality experiment. The drawback is that subjective study is usually expensive and not always reliable.

In [4], the authors present an objective measurement based on the Peak signal-to-noise ratio PSNR metric and salience maps to assess areas that catch user attention in a virtual reality environment.

In this paper we present our investigation of the use of ,view synthesis quality assessment (VSQA) [1], a metric that was originally designed for assessing novel synthesis using processes such as panorama stitching.According to the authors,

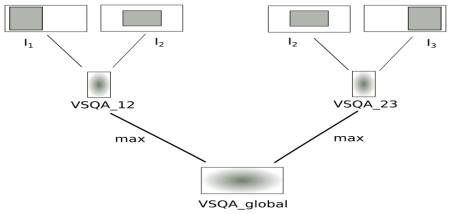

Fig. 2. A simplified figure for the construction of our error detection on a panoramic video frame.

the human vision system (HVS) is mostly sensitive to local image variations in texture, gradient orientation diversity and high contrast areas. Therefore, they extend the well-established image metric structural similarity image index (SSIM)[5] with three visibility weighting maps, that increase or decrease the distortion value depending on its visual saliency. Further details of our implementation are provided in the next section.

III. Proposed approach

Our approach is to provide a quality evaluation for the panorama before the actual blending takes place. Performing error calculation prior to blending provide three advantages: first, although blending strives to remove some artifacts, it is a blind method that can introduce new artifacts by removing parts of objects or mistakenly erasing something that is not actually an error. Second, once images have been blended into the final panorama, it is very difficult to recover the original images, which are as the name of the method implies, blended and mixed together in the overlapping areas, therefore post-processing to correct defects will also be difficult. Finally, to detect misalignment and discontinuities, it is essential to compare the structural dissimilarities between intersecting views, which is only available prior to blending. As explained in the previous sections, we use the VSQA quality metric [1] calculated as in 1, which was designed for novel view synthesis to calculate our error map.

$$VSQA(i,j) = \text{dist}(i,j).[W_t]^{\delta}.[W_o]^{\epsilon}.[W_c]^{\zeta}. \quad (1)$$

where $dist$ is the chosen metric, in this case SSIM [5], calculated between a reference view and a synthesized view. This metric is weighted by 3 maps, each representing a type of local feature to which the human eye is most sensitive. Details are found in [1].

Our pipeline to produce the global error map proceeds as follows:
Consider N views at a time t, after calculating pairwise matches $P_n(i,j)$, for each pair I_i and I_j, we calculate the region of overlap $I_i \cap I_j$ and we compute VSQA metric between the region of interest in each view δI_i and δI_j.
We finally calculate the equation 2 to generate a global map for the whole panorama by taking the maximum contributing pixel in all overlapping views as shown in figure 2.

$$VSQA_{global(}i,j) = max_{i,j} VSQA_{i,j}(\delta I_i, \delta I_j). \quad (2)$$

Where i,j represent pixel location.
In order to properly assess video frames that are produced by algorithms such as that of Perazzi et al. [6], we don't compare the two views directly. Instead, we calculate the motion field from one view to the other, we then do a back-ward warp from the source view to the target and we calculate our metric according to equation 3. We do this to represent parallax compensation as presented in [6].

$$VSQA_{global}(i,j) = max_{i,j} VSQA_{i,j}(\delta I_i, warp(\delta I_j)). \quad (3)$$

All of the calculations above take place before blending the different views together. As mentioned earlier, the blending step aims mainly to remove as many of these errors as possible, though it does not succeed in all the cases. The multi-band blend described in [7] usually uses a Voronoi mask that chooses the blending line regardless of the image content. Therefore, the pixels on and around this boundary line usually have the most visible defects and they tend to disappear the farther we move far away from it. Based on this fact, we propose to create a weighting mask around this blending edge, which will give more weight to the pixels that fall onto and around this line and decreases gradually the more we go farther away. Within the same iterations over pair-wise matches as described in the previous sub-section, for a pair of views I_i and I_j, we calculate the Voronoi seam cut which produces a mask for each view M_i and M_j that determine the cutting line between both views. We are also interested only in the region of intersection between the two images, so we use the sub-masks δM_i and δM_j. In order to create the desired mask, we calculate a distance transform from that line for each of the latter sub-masks, we then calculate a common mask that will be applied to the resulting VSQA as the OR between δM_i and δM_j and we get a mask M_{blend} that we normalize afterwards. We multiply this mask to our VSQA computed at each step in order to enforce errors at the region where the transition between images takes place and attenuate errors farther away from this boundary as described in equation 4. We call this measure MVSQA.

$$MVSQA = M_{blend}.VSQA. \quad (4)$$

We generate the global MVSQA with the same process used to calculate the composite VSQA as described previously.

IV. Experimental results

In order to test our method, we used the dataset *Opera*, a video sequence taken by a 5 GoPro camera-rig, provided by Perazzi [6] for their work on panoramic videos. We apply equation 3 to represent stitching methods incorporating parallax compensation. We also took our own panoramic video using a 3-camera rig designed by [9] formed of 3 Panasonic GH2 cameras with 20mm lens each. Video frames were generated using the open source software Hugin [8] for panorama creation, with graph-cut multi-band blending, for which we used equation 2. The results shown in 3 show a promising prediction for zones of potential errors not

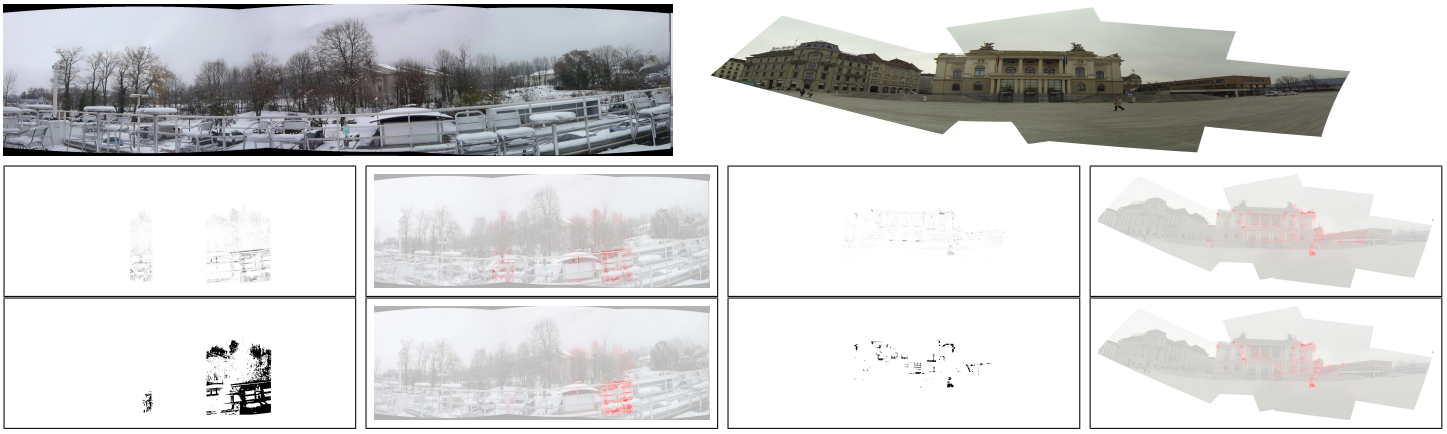

Fig. 3. The top row shows a panoramic scene of snow created by Hugin [8] and an opera building scene created by the method in [6]. The second row shows the results as applied in equation 2 for the left panorama and 3 for the right. Finally the last row is the results of applying equation 4 and thresholding it foe better visibility.

only spatially but across the whole sequence. Repeating the process for some key-frames in the video, can show which errors persist and which appear sporadically. It can also be noticed that the error seems concentrated in the right middle part of the panorama in the dataset *Opera* which contains four out of the five views overlapping. The suggested mask permitted to focus on errors around the blend mask and therefore reducing the number of false positives.

In order to obtain a metric index out of our distortion map, we calculated a score according to [1], which consists in counting the number of remaining erroneous pixels after applying a threshold. We used the same method of spatial pooling to compare our results with basic SSIM. Table I shows the resulting scores in percent of remaining pixels for VSQA and MVSQA described in equations 2 and 4 as well as basic SSIM [5]. The measures were applied to 3 chosen frames where we could see clear parallax errors. Figure 4 shows examples of parallax erros that appeared after stitching and their corresponding maps VSQA, MVSQA and SSIM.

Fig. 4. Zoom on error in Opera sequence. Top row is panorama at time t=105, the one below is t=385.Then from left to right, the figure shows the original view before stitching,then the image after being stitched. Then the error maps for VSQA, MVSQA and SSIM respectively are shown.

Metric / Score in %	at t=84	at t=105	at t=385
MVSQA	0.38	0.55	0.26
VSQA	0.56	1.04	0.29
SSIM	9.63	10.58	8.45

TABLE I
RESULTS OF SPATIAL POOLING

V. CONCLUSION

We presented a metric for panoramic video quality assessment integrated within the stitching process. Our experiments show it can be beneficial to compare images before blending them all together, as it can show potential artifacts locations. The application of the described blend mask filters the errors further, which accentuates those who can persist after blending. We continue to work on this approach in the goal of adding a temporal factor that will help to assess a video globally using motion estimation. Final panorama should also be taken into account in the calculations. We will be conducting a user study to obtain subjective quality measurements that can help us validate our method as well as comparing it to other approaches.

REFERENCES

[1] P. Conze, P. Robert, and L. Morin, "Objective View Synthesis Quality Assessment," in *Stereoscopic Displays and Applications*, ser. Proc SPIE, SPIE, Ed., vol. 8288, San Francisco, United States, Jan. 2012, pp. 8288–56.

[2] L. Yang, Z. Tan, Z. Huang, and G. Cheung, "A content-aware metric for stitched panoramic image quality assessment," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, "Subjective and objective quality assessment of panoramic videos in virtual reality environments," *2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, vol. 00, pp. 163–168, 2017.

[4] M. Xu, C. Li, Z. Wang, and Z. Chen, "Visual Quality Assessment of Panoramic Video," *ArXiv e-prints*, Sep. 2017.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[6] F. Perazzi, A. Sorkine-Hornung, H. Zimmer, P. Kaufmann, O. Wang, S. Watson, and M. H. Gross, "Panoramic video from unstructured camera arrays," *Comput Graph Forum*, vol. 34, no. 2, 2015.

[7] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Trans. Graph.*, vol. 2, no. 4, pp. 217–236, Oct. 1983. [Online]. Available: http://doi.acm.org/10.1145/245.247

[8] H. P. photo stitcher, "Open source software," http://hugin.sourceforge.net/.

[9] "Amiqual4home," https://amiqual4home.inria.fr/.