Sports VR Content Generation from Regular Camera Feeds

Kiana Calagari[1], Mohamed Elgharib[2], Shervin Shirmohammadi[3], Mohamed Hefeeda[1]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
[3]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

ABSTRACT

With the recent availability of commodity Virtual Reality (VR) products, immersive video content is receiving a significant interest. However, producing high-quality VR content often requires upgrading the entire production pipeline, which is costly and time-consuming. In this work, we propose using video feeds from regular broadcasting cameras to generate immersive content. We utilize the motion of the main camera to generate a wide-angle panorama. Using various techniques, we remove the parallax and align all video feeds. We then overlay parts from each video feed on the main panorama using Poisson blending. We examined our technique on various sports including basketball, ice hockey and volleyball. Subjective studies show that most participants rated their immersive experience when viewing our generated content between Good to Excellent. In addition, most participants rated their sense of presence to be similar to ground-truth content captured using a GoPro Omni 360 camera rig.

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → **Virtual reality**;

KEYWORDS

2D-to-VR conversion, 360 videos, virtual reality

1 INTRODUCTION

Virtual reality is currently experiencing a growing interest in the multimedia industry. Despite large investments from giants such as Facebook, Google, Microsoft, Samsung, and others, one problem remains that prevents VR from being adopted on a wider scale; the lack of real content. The vast majority of current content is synthetic, generated for the gaming industry. The only approach for generating real content is by using VR capturing devices. Such devices, commonly referred to as VR camera rigs, contain multiple cameras stacked next to each other in a way that maximizes the field of view [2, 11, 33]. Camera outputs are then stitched together to enhance the overall sense of immersion.

Current solutions for capturing high-quality VR content require upgrading the entire production pipeline. Such upgrade is expensive to set-up and operate, which makes it an unappealing solution. We propose an alternative approach for VR content generation that converts the traditional broadcast to VR through post-processing.

Most large sporting stadiums contain multiple high-end cameras, capturing the field from different positions. These cameras are operated by a professional filming staff that can create production quality content. For instance, broadcasting a FIFA World Cup game often involves more than 20 cameras [13]. These cameras capture the field from different angles and different positions, including a few main cameras positioned on the halfway line, a high left and a high right camera. The report in [34] shows the most common camera positions for different field sports, including basketball, and ice hockey. In most games, there is at least one main camera positioned usually in the middle of the field. This camera follows the action with a wide angle covering around 50% of the field. With such setup already in place, we propose a solution for high-quality VR content generation. Our solution utilizes existing video feeds in a post-processing step without the need of upgrading the entire production pipeline.

Producing VR content from the traditional video feeds is a quite complex task and requires addressing multiple challenges. First, we need to widen the field of view to at least 180 degrees. We achieve this by utilizing the motion of the main camera to generate a wide-angle static panorama. The field area is then overlaid on the panorama using Poisson blending to allow seamless blending. Second, due to the limited coverage of the main camera, players tend to appear/disappear throughout the recording. Such effect significantly degrades the feeling of immersion. To overcome this problem, we identify and retrieve the missing players from different camera feeds. For this, all feeds need to be aligned with respect to a reference frame, which is challenging because of the large distance between cameras. This distance causes the position of objects and the orientation of lines to appear differently when viewed from different cameras. This effect is referred to as parallax. Large amounts of parallax cause state-of-the-art image alignment techniques, e.g. [41], to fail. To address this problem, we remove parallax by first obtaining camera parameters, and estimating the 3D position of each pixel. We then warp each feed to the position of the main camera.

To evaluate our method, we captured sports games using 3 regular cameras and a 360 camera simultaneously. Using our technique, we generated VR content from the 3 regular video feeds, which were positioned on the left, middle, and right side of the field. We conducted subjective studies in which participants were asked to rate their sense of presence and perceived video quality when viewing our generated content. In addition, they were asked to compare our content with content captured using the 360 camera. Our results show that most participants rated their sense of presence

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM'17, October 23–27, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3123266.3123315

between Good to Excellent. Also, our generated content was rated similar to the captured 360. Analysis on missing players show that our method retrieves missing players accurately with a maximum displacement error of around 20 *cm*.

The rest of this paper is organized as follows. Section 2 provides a summary of the related work. Section 3 describes our proposed method in details. Section 4 presents an extensive evaluation of our technique, and Section 5 concludes the paper.

2 RELATED WORK

Recently, VR has gained strong interest from both the industrial and research communities. Stengel et al. [35] proposed an affordable solution for VR headsets which incorporates eye tracking and tackles motion sickness. Perrin et al. [31] addressed quality assessment of VR content through a multi-modal dataset, while Chang et al. [9] proposed a methodology for quantifying the performance of commodity VR systems. Zare et al. [42] presented a streaming solution for VR that can reduce the bitrate down by 40%. A survey on VR streaming solutions is presented in [10]. While this line of research addresses a wide range of VR topics, they do not address the content generation step.

Capturing VR content requires a camera rig with a field of view of 180 to 360 degrees. A number of such rigs have been recently introduced, including GoPro's Omni [19], Samsung's Beyond [33], Facebook's Surround 360 [11] and Google's Odyssey [2]. These rigs contain multiple cameras stacked next to each other in a way that maximizes the field of view. Various software tools are used to allow seamless stitching and blending of the different camera views. Specialized filming teams are needed to operate the production pipeline. Companies offering such solutions include NextVR [28] and Jaunt [23]. However, using any of these solutions requires upgrading the entire production pipeline. This is often an expensive and inconvenient process.

VR content generation is based on creating a wide-view panorama of the scene of interest. One approach for panorama generation is through image stitching . Here, images are aligned in space by estimating a warping field through feature point matching. Many techniques assume simple camera rotations [4, 27, 39] and/or planar scenes [1]. Others relax these constraints through dual homography [16] or by smoothly varying the affine/homography [8, 24, 41]. Zaragoza et al. [41] rely on projective transformation and estimate local homography for alignment. Chang et al. [8] use a parametric warp which is a spatial combination of a projective transformation and a similarity transformation. Perazzi et al. [29] use patch based error metric similar to optical flow to estimate warping.

Most current stitching techniques allow only a small parallax and hence assume images are taken from nearby cameras. Recently, Zhang et al. [43] proposed an approach that relaxes this assumption. A mesh-based framework is proposed that optimizes for both alignment and 2D regularity. Interesting results are generated that can handle moderate parallax and moderate deviation from planar structures. However, limitations still exist, such as the inability of handling straight lines across multiple cameras. Line straightness can only be preserved locally in each image, but not across images. Such artifacts are problematic for sports content, as it is often crucial to preserve the straightness of field lines.

Another approach for generating panoramic images is through 3D modeling and texture mapping. Multiview techniques perform 3D model reconstruction by estimating point locations via feature correspondence. VisualSFM [40] provides a GPU based optimized implementation of such techniques. Generating dense 3D reconstruction in outdoor environment is still a challenge. While the technique in [22] produces good results with large datasets, the reconstruction quality highly depends on good feature point correspondence. Such correspondence is not necessarily available in sports data due to the low textured nature of playing fields.

Sports content has special properties. Specifically, the presence of all players and the straightness of the lines are of major importance, while the background is less of a concern. The main component of sports VR content is a wide-view panorama with all players present at every time instant. Fehn et al. [12] use two nearby high resolution cameras to generate a high resolution (5k) panorama of a soccer match. The two cameras are planted in a way to have a full coverage of the field. Similarly, Stensland et al. [36] use a camera array of four nearby cameras to generate a panoramic video as part of a sport analysis system called Bagadus. These deployments, however, require a special setup which is difficult to achieve with the current broadcast systems.

Inputs from multiple cameras have been used to create free view point videos (FVV) [3, 17, 20]. In FVV, the task is synthesizing novel views from the available ones. This process, however, contains a number of stages such as camera calibration, segmentation, depth estimation and 3D reconstruction. Multiple works, e.g. [21, 26], proposed a number of FVV approaches for soccer. However, they all require pre-calibrated static cameras. Such set-up is hard to satisfy in sporting events with highly dynamic nature. Germann et al. [17] presented a FVV technique that robustly handles multi-cameras with wide baselines in an uncontrolled environment. Feature correspondence between cameras are found and back-projection errors are used to estimate a novel scene reconstruction method. Angehrn et al. [3] acknowledged that aligning multiple cameras is one of the most challenging tasks for FVV. To improve the performance of this step, they introduced the concept of a static master camera. All cameras are aligned to this camera.

A wide-view panorama can also be generated using different time instants of a single camera. This requires aligning each video frame to a reference panorama plane. Ghanem et al. [18] proposed matching global features such as image patches rather than matching small salient points. Their approach, however, does not consider the temporal stability of the estimated alignments and hence may generate shaky results. Carr et al. [7] proposed a gradient-based alignment to edge images. Due to computational complexity of the approach, the calibration does not scale well with video.

Summary/Motivation: Despite the rich literature of image stitching and panorama generation techniques, up to our knowledge there is no prior work on producing VR content from the traditional broadcast pipeline. We present a computational approach for doing so and we tailor our solution for sports. Our solution exploits the movement of a main camera to generate a wide-view panorama and utilizes other cameras to estimate missing data such as players. We optimize the visual quality of our results using careful image blending and accurate alignment suitable for our problem.

3 PROPOSED METHOD

3.1 Overview

The output of our proposed technique is a 360 video in equirectangular format that can be viewed on VR headsets or regular displays using 360 video players. In order to generate this 360 video from regular camera feeds, one of the cameras is chosen as the user's viewing position. We refer to this camera as the main camera. While the best choice would be a wide-view camera that follows the action, any rotating camera can be sufficient for our method. Note that with multiple choices for the main camera, multiple VR videos can be generated from different positions and angles, providing the user with multiple options. For any chosen main camera, all cameras within its 180 degrees field of view can be used as complementary feeds and help in covering the whole field.

Our technique consists of three stages (Fig. 1). In the first stage, we use the main video feed to generate a wide-view static panorama by means of image registration and median filtering. This panorama is used as the background of our 360 environment, in which the field and players will be then overlaid on. Although the background remains static, subjective studies show that it has little impact on the sense of presence, as it is not the focus of attention. In the second stage, we remove parallax between all video feeds by warping them to the position of the main camera. In the third stage, we use Poisson blending to first overlay the main feed on the panorama, and then copy the missing players from the complementary feeds. In the following sections, we describe each stage in more detail.

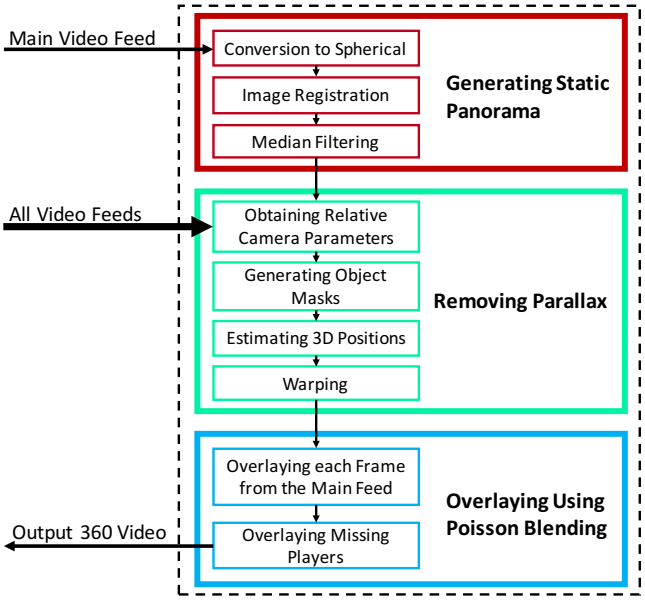

Figure 1: The 3 stages of our immersive content generation technique, and their main components.

3.2 Generating Static Panorama

The viewing angle in regular sports videos is usually not wide enough for an immersive experience. In order to improve the sense of presence, a wider viewing angle is needed. As a result, we increase the viewing angle by utilizing the camera rotation, and generating a panorama image which includes the static parts of the scene. This stage can be performed only once, or periodically during a long game to capture any changes in the background. Only the main video feed is used in this stage. It is recommended to use a shot in which the camera rotates over a large angle and with minimum zoom. This widens the viewing angle of the generated panorama. In order to display a viewing angle greater than 180 degrees, we perform a planar to spherical conversion on each frame prior to any processing. The camera rotation is then transformed to a wider viewing angle by aligning the spherical frames using image registration techniques, and applying median filtering. Fig. 2 shows an example static panorama generated from a basketball game.

Conversion to Spherical: Aligning planar images using projective transformation can cause shape and size distortion. This problem becomes more severe as the angle between the frames increases. While methods such as [8] try to overcome this problem, a viewing angle above 180 degrees cannot be achieved in planar format. In order to produce a panorama image with a viewing angle above 180 degrees from planar video frames, we convert the frames to spherical projection. The equirectangular projection is a standard way of projecting the 3D world onto a flattened sphere. It is an image with size $2\pi r \times \pi r$ that will be wrapped around a sphere when viewed in 360 degree. Note that r is constant, and can be chosen arbitrarily based on the desired output size (resolution). To map a frame to an equirectangular projected image, for each pixel in the frame (x, y), the spherical coordinates (θ, φ) are calculated using Eq. (1). Here, the origin is taken at the centre of the frame. α is the camera angle. Z_{img} is the distance of the frame from the camera and is a function of the focal length. The pixel is then mapped to $(r\varphi, r\theta)$ in the equirectangular image.

$$\varphi = \arctan\left(\frac{x}{Z_{img}\cos(\alpha) + y\sin(\alpha)}\right),$$
$$\theta = \frac{\pi}{2} + \arctan\left(\frac{Z_{img}\sin(\alpha) - y\cos(\alpha)}{\sqrt{(Z_{img}\cos(\alpha) + y\sin(\alpha))^2 + x^2}}\right). \quad (1)$$

Image Registration: By performing image registration on the equirectangular frames, we transform the camera rotation to a wider angle of view. We automatically perform registration using feature based image registration techniques such as [4]. First, we extract and match SIFT features [25] between consecutive frames. Using RANSAC (random sample consensus) [14] we select a set of inliners that are compatible with a homography between the frames. We then align the frames according to the homography by applying a similarity transformation.

Median Filtering: The static panorama is extracted from the aligned frames using median filtering. We assign the value of each pixel in the panorama to be the median across all aligned frames. By applying median filtering, the moving objects will be removed, leaving only the static background. Note that applying median filter only on key frames can generate sharper results than applying it to all frames.

Figure 2: Example of a static panorama generated from a basketball game.

3.3 Removing Parallax

In a regular sports production the cameras are usually placed meters away from each other, causing a huge amount of parallax between them. Fig. 3 shows an example of 3 frames from cameras in different positions captured at the same time instant. Notice how the parallax affects line orientations and player positionings. In this section, we describe how we warp the camera feeds to remove such parallax.

Obtaining Relative Camera Parameters: In order to remove the parallax, we warp all feeds to the position of the main camera. We do so using relative camera positions and the 3D position of each pixel in space. While the availability of such information is desirable and can further improve the results of our technique, it is more practical not to always assume such information is given. Thus, we obtain an estimation of the relative camera parameters using the VisualSFM [40] software, and estimate the 3D position of each pixel using plane fitting and object masks. VisualSFM performs a sparse 3D model reconstruction using structure from motion techniques and provides an estimation of the relative camera positions (C), relative camera rotation matrixes (R), camera focal lengths (f), and sample 3D points (X_w, Y_w, Z_w).

Generating Object Masks: Object masks are used throughout our technique in different stages and for multiple purposes. They are required for estimating the 3D position of the players, identifying the missing players, and copying them on the panorama. An object mask indicates the pixels which are not part of the static background. For sports, this is mainly the players. To create these masks we use the background subtraction technique. We first apply a median filter on every group of frames to get the background. Choosing a larger group size can better identify slower moving objects, such as a player that is still for a few seconds, but it may introduce more noise. The moving objects are then detected by subtracting each frame from its background. We further enhance the object masks by applying several morphological filters. Fig. 4(a) shows the generated object mask for the frame shown in Fig. 3(c).

Estimating the 3D Position of Pixels: When capturing an image, a 3D point (X_w, Y_w, Z_w) in world coordinates is first projected to the camera coordinates (X_c, Y_c, Z_c) through Eq. (2). In this equation, R is the rotation matrix of the camera, and T is the translation vector calculated based on the camera positions ($T = -RC$). The 3D point is then projected on the 2D image through Eq. (3), where (x_i, y_i) represent the image coordinates with the origin being at the image centre. Due to the loss of the third dimension in Eq. (3), this projection is non-revertible unless Z_c is known. As a result, to find the 3D position of a pixel, we should estimate its Z_c.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} R|T \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} fX_c/Z_c \\ fY_c/Z_c \end{bmatrix} \tag{3}$$

While the camera can have different tilts, its x axis is usually set parallel to the ground. Thus, in the camera's coordinate system, the field is a plane parallel to the x axis which can be presented as $bY_c + cZ_c = 1$. Since usually a large area of the frame is covered by the field, we estimate the plane parameters b and c by fitting a plane to sample (X_c, Y_c, Z_c) points. To obtain such samples, we use Eq. (2) to project the sample 3D points provided by VisualSFM to the camera coordinates. With the plane parameters in hand, Z_c of each field pixel (x_i, y_i) can be estimated through Eq. (4). From the non-field areas, our main concern is the players, which are indicated by the object mask. Based on the object mask, we estimate the Z_c of each player pixel to be the Z_c of the place where the player's feet touch the ground. Note that while sophisticated depth estimation techniques, such as [6], can be used for estimating Z_c, they are not necessary since, as shown in Sec. 4.4, the current method achieves a fine warping accuracy and the error is fairly small.

$$Z_c = \frac{1}{by_i + c}. \tag{4}$$

Warping: Using the 3D pixels positions and the relative camera parameters, we warp each video feed and its corresponding object mask to the position of the main camera. To do so, for each pixel (x_i, y_i), we first revert the camera projection to find the world coordinates. We then project each point back to a 2D image, where the new coordinates $(x_{i_{main}}, y_{i_{main}})$ are calculated according to the main camera parameters. This process is shown in Eq. (5) and Eq. (6). In Eq. (5), an accurate estimation of the relative camera positions C, which manifest itself in the translation vectors T, can successfully remove all parallax. Note that, as a result of such warping, parts of the field that were originally occluded by the players may now become visible, causing empty shadow-like holes under each player. Such holes can be filled by inpainting techniques. In our experiments, we use an averaging approach for filling such holes. Fig. 4(b) shows an example of our warp applied to Fig. 3(c). Notice the similarity between the warped image and the main feed (Fig. 3(b)) in the field lines orientations and positioning of the players.

(a) Left camera (b) Middle/Main camera (c) Right camera

Figure 3: Sample frames of 3 different video feeds positioned around the field. All frames are shot at the same time instant.

(a) Object mask for right view (b) Right view warped to the position of the main camera

Figure 4: Removing parallax between the right camera frame in Fig.3(c) and the frame from the main camera in Fig.3(b) using its object mask. The object mask is generated using the background subtraction technique.

$$\begin{bmatrix} X_{c_{main}} \\ Y_{c_{main}} \\ Z_{c_{main}} \end{bmatrix} = R_{main}(R^{-1}(\frac{Z_c}{f}\begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix} - T)) + T_{main} \tag{5}$$

$$\begin{bmatrix} x_{i_{main}} \\ y_{i_{main}} \end{bmatrix} = \begin{bmatrix} f_{main} X_{c_{main}}/Z_{c_{main}} \\ f_{main} Y_{c_{main}}/Z_{c_{main}} \end{bmatrix} \tag{6}$$

3.4 Overlaying Using Poisson Blending

In order to generate our output 360 video, we need to overlay parts of each video feed on the static panorama. To seamlessly blend the copied parts with the background, we use Poisson image editing. Poisson image editing is known as a seamless image cloning algorithm based on gradient field [30] and it produces more plausible results than just simply overlaying the objects on the panorama. However, a limitation of the Poisson image editing approach is that the color of the source image gets totally adapted to the background image. To overcome this problem, we utilize an image cloning algorithm based on a modified Poisson problem [38]. The modified version has a color preserving parameter. A large color preserving parameter perfectly preserves the color of the source and background in the overlaid result.

For each frame, we first overlay the main feed. Players missing from the main feed are then identified and copied from the complementary feeds. A main feed that follows the action is most likely to cover the ball and most players, leaving only a few players missing.

Overlaying Frames from the Main Feed: To overlay a frame from the main feed on the panorama, we first convert it to spherical format using Eq. (1). We then perform image registration by matching SIFT feature points between the frame and the panorama, and using RANSAC to select a set of inliers. After aligning the frame with the panorama by applying a similarity transformation, we seamlessly blend the frame borders by means of Poisson blending. Note that in order to reduce the effect of possible misalignments, if a player is on the borders of the frame, it is removed and considered as a missing player. Fig. 5(a) shows a sample frame from the main feed (Fig. 3(b)) overlaid on the static panorama.

Overlaying Missing Players: Missing players are identified using our object masks. If an object in the mask is partially or completely outside the area covered by the main feed, it is considered missing and is copied. For example, after warping the right view object mask in Fig. 4(a) to the position of the main camera (similar to Fig. 4(b)), 2 of its objects fall into the area covered by the main feed (Fig. 5(a)). The other 3 objects, however, are identified as missing.

Similar to the main feed, for identifying and copying the missing players, we should first align the warped complementary frames with the panorama. With the parallax removed, this alignment can be performed rather accurately. However, for a better alignment, and to overcome possible errors in estimating the camera parameters (Sec. 3.3), we perform image registration on planar images. To do so, we keep the warped frames in planar format and convert the field area of the panorama from spherical to planar format using Eq. (7). We then use SIFT and RANSAC to calculate the homography, and align images by applying a projective transformation. Missing players are then identified, copied, and blended seamlessly with the panorama using Poisson blending.

Finally, after all missing players have been overlaid, the planar panorama is converted back to spherical projection and placed in its corresponding location in the 360 panorama. Fig. 5(b) shows a zoomed-in version of a final 360 frame after all missing players (blue arrows) have been overlaid.

(a) Main view overlaid on the static panorama

(b) Zoomed version of final panorama with all missing players overlaid

Figure 5: Overlaying the main feed on the static panorama, and copying the missing players from the left and right feeds. The blue arrows indicate that the players are copied from the left or right feeds.

$$x = \tan(\varphi)(Z_{img}\cos(\alpha) + y\sin(\alpha)),$$

$$y = Z_{img}\frac{\sin(\alpha) - \frac{\tan(\theta-\frac{\pi}{2})\cos(\alpha)}{cos(\varphi)}}{\cos(\alpha) + \frac{\tan(\theta-\frac{\pi}{2})\sin(\alpha)}{cos(\varphi)}}. \quad (7)$$

4 EVALUATION

To evaluate our VR content generation technique, we conduct subjective studies to measure the average subject satisfaction when observing our generated content. We also compare our results against content captured using 360 camera. In addition, we analyze the accuracy of our technique in retrieving missing players.

Our technique requires sports video feeds from different cameras around the field. At least one of the cameras needs to be moving. While such setup is realistic for broadcasting companies [34], we do not have access to their captured feeds. In addition, to the best of our knowledge, all available datasets such as [15, 32, 37] only provide feeds from static cameras. Hence, we captured our own data, while simulating broadcasting setups.

4.1 Setup

We captured multiple games using a GoPro Omni 360 camera as well as 3 individual GoPro Hero4-black cameras. All cameras captured the scene simultaneously. The Omni camera rig consists of 6 GoPro Hero4-black cameras, capturing in different directions. It was deployed in the middle of the field to capture 360 content. We treat this captured 360 content as ground-truth and compare it against our own reconstruction. The 3 individual cameras were deployed at the left, right and middle of the field, capturing the scene in 4K resolution. We synced the 3 individual cameras by pairing them with the GoPro Wi-Fi remote. The synchronization was further refined manually. The middle camera rotates with the action, and is considered as the main feed. The left and right cameras are static. Initially, the middle camera is rotated with a wide angle so it would capture most of its surrounding and cover around a 360 degrees. GoPro Hero4-black cameras are wide-angled and do not provide zooming options. Hence, to simulate professional content more accurately, we zoom on our 3 individual camera feeds in a post-processing step. For this we use the GoPro Studio software.

While our technique can be used for all field sports, we used data from 3 different games: basketball, ice hockey, and volleyball (Fig. 6). From each game, we chose a 30-second sequence and converted it to VR content using our technique. For the same sequence, we also stitched the GoPro Omni feeds using its recommended software (Autopano Video) to create what we refer to as the original 360.

In our subjective experiments we assess sense of presence and video quality for both the original 360 and our generated content. A high sense of presence means that the participants are fully immersed into the action. For video quality, we focus on the amount of artifacts. Generating VR content relies heavily on image processing techniques and is therefore prone to various artifacts. By assessing the quality we measure the amount and visibility of such artifacts.

4.2 Evaluation of our Technique

We conduct a subjective study to measure the average subject satisfaction when viewing our generated content. Fifteen participants

Figure 6: Examples of final panoramas generated by our technique for different games: basketball (top), hockey (middle), and volleyball (bottom). The blue arrows indicate the players that have been copied from the left or right feeds.

took part in our experiments. They were all computer science students and researchers. We used Oculus Rift to display the VR content. We displayed the games in random order. Prior to the actual experiments, we showed the participants samples of professionally produced 360 videos from the Rio olympics. This familiarized the participants with the VR device and the 360 environment and hence stabilized their expectations. We noticed that participants tend to move their head more when they first wear the device, and focus more on the games as they get used to the experience.

We used the standard ITU continuous scale [5] to rate both video quality and sense of presence. The labels marked on the continuous scale are Excellent, Good, Fair, Poor, and Bad. We asked participants to mark their scores on the continuous scales. Their marks were then mapped to integer values between 0-100, and averaged to calculate the mean opinion score (MOS). Participants were asked to clarify all their questions and ensure their full understanding of the experimental procedure.

Fig. 7 shows the MOS for different games. Error bars represent the standard deviation. Most participants rated both video quality and sense of presence in the range of Good to Excellent for all games. This means that they were well immersed in the 360 experience. Between the three games, hockey has the least score. This is because the low-textured hockey field makes it difficult to perform accurate feature matching and alignment.

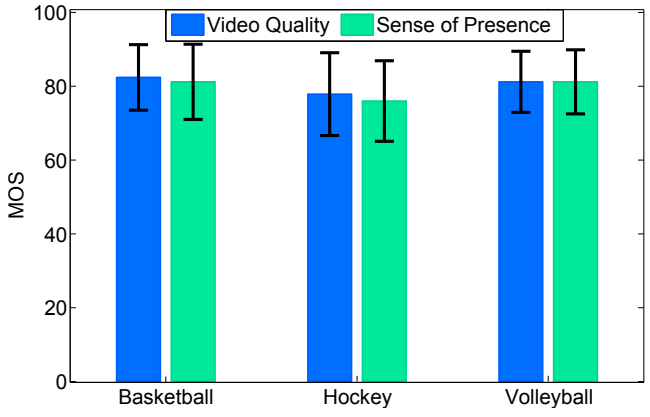

Figure 7: Mean opinion scores (MOS) of video quality and sense of presence for different games.

In addition, after the experiment, we asked the participants whether they noticed that the background was static. While some participants didn't notice it at all, as they were focused on the field and players, the ones that did, stated that it had affected their sense of presence marginally. Note that in our technique the background can change periodically at the expense of more computational cost.

4.3 Comparison against Original 360 Content

We compare the results of our technique against original 360 content captured at the same time instant using a 360 camera. For this experiment, we use the double stimulus method (DSCQS) [5], where participants view both content in random order and can re-view them as many times as they need. Participants were asked to rate the video quality and sense of presence for both content using the standard ITU continuous scale. Their marks were then mapped to integer values between 0-100. We calculated the mean of difference opinion scores (DMOS) by averaging the difference opinion scores (= score for our technique - score for original 360). A DMOS of zero implies that the results of our technique are judged the same as the original 360, while a negative DMOS implies that our result has a lower quality/sense of presence than the original 360.

Fig. 8 shows the DMOS of both video quality and sense of presence for each game. Error bars represent the standard deviation. The small DMOS values indicate that most participants found their immersive experience to be quite the same when comparing our generated content against the ground-truth content captured using 360 camera. In addition, the only statistically significant difference reported (p-value < 0.05) is the sense of presence for hockey.

4.4 Analysis of Player Placements

Copying the missing players is an important aspect of our technique. Failing to accurately place the players at their correct locations can cause sudden player movements that may seem unnatural and disturbing. Fig. 6 shows examples of final panoramas generated by our technique. Note that in these examples some feeds were zoomed more than usual, in order to have more missing players for demonstration purposes. The blue arrows indicate the missing players that were copied from the left and right feeds. To analyze the effectiveness of our technique in retrieving these missing players, we measure the amount of their displacement. We use the originally captured wide-angle main feed as reference. We measure the distance between the position of each copied player and its original position in the reference frame. We define the position of a player, as the pixel coordinates of the place where the player's feet touch the ground. Fig. 9 shows the average displacement. Error bars represent the standard deviation. It can be seen that the displacement is highest for hockey, with a maximum around 10 pixels. This is because hockey is more prone to misalignment errors due to its low-textured field with high intensity color. However, we should note that a displacement of 10 pixels in the panorama translates to a distance around 20 *cm* in a real field, which is fairly small.

5 CONCLUSIONS AND FUTURE WORK

We presented a technique for generating VR content for sports from common broadcast camera feeds. While current solutions for producing high-quality VR content require upgrading the entire production pipeline, our technique utilizes the existing camera setup to generate immersive content. We assume the presence of at least one camera with rotational movement and two or more complementary cameras which altogether cover the whole field. Our method has three main stages: (1) creating a wide-angle panorama, (2) removing parallax and aligning all video feeds, and (3) overlaying the field and the missing players on the panorama by means of

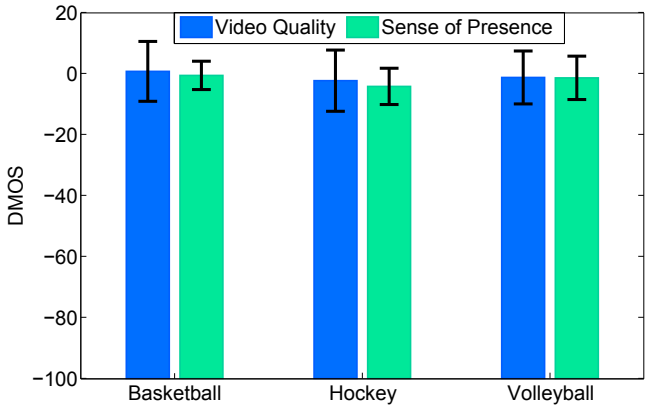

Figure 8: Difference mean opinion score (DMOS) between content generated using our technique and original 360 content captured using 360 camera. A value of zero implies that the results of our technique are the same as the original 360.

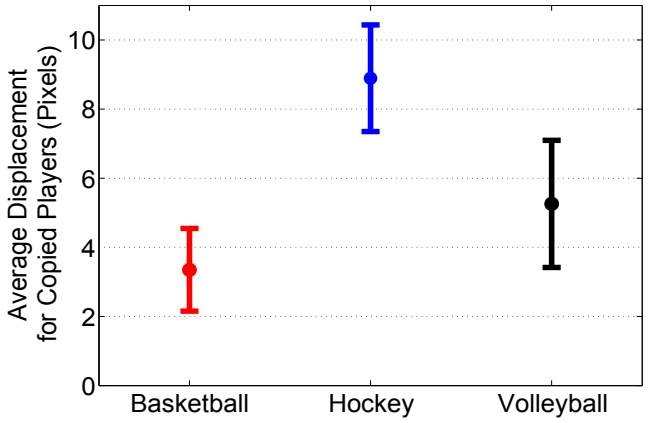

Figure 9: Average displacement of copied players for different games. Error bars represent the standard deviation.

Poisson blending. Subjective experiments show that our results are comparable with the original 360 content in terms of both quality and sense of presence. In addition, MOS ratings indicate that most participants experienced a strong sense of immersion.

Future work can address better handling of players in cluttered regions. Current results can be temporally inconsistent in such cases. The results can be further improved by incorporating depth information from infrared cameras as well as other auxiliary data, e.g., ground-truth camera positions. In addition, although the static background didn't damage the immersive experience and only marginally affected the sense of presence, exploring ways of capturing some of the fans dynamics and including it in the panorama without too much complexities may enhance the results.

6 ACKNOWLEDGMENTS

This work is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by the Qatar National Research Fund (grant # [NPRP8-519-1-108]).

REFERENCES

[1] Aseem Agarwala, Maneesh Agrawala, Michael Cohen, David Salesin, and Richard Szeliski. 2006. Photographing Long Scenes with Multi-viewpoint Panoramas. *ACM Transactions on Graphics (TOG'06)* 25, 3 (2006), 853–861.

[2] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. 2016. Jump: virtual reality video. *ACM Transactions on Graphics (TOG'16)* 35, 6 (2016), 198.

[3] Florian Angehrn, Oliver Wang, Yağiz Aksoy, Markus Gross, and Aljoša Smolić. 2014. MasterCam FVV: Robust registration of multiview sports video to a static high-resolution master camera for free viewpoint video. In *Proc. of the International Conference on Image Processing (ICIP'14)*. Paris, France, 3474–3478.

[4] Matthew Brown and David G Lowe. 2007. Automatic panoramic image stitching using invariant features. *International journal of computer vision* 74, 1 (2007), 59–73.

[5] ITU-R BT.500-13. 2012. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland.

[6] Kiana Calagari, Mohamed Elgharib, Piotr Didyk, Alexandre Kaspar, Wojciech Matusik, and Mohamed Hefeeda. 2015. Gradient-based 2D-to-3D Conversion for Soccer Videos. In *Proc. of the ACM Multimedia Conference (MM'15)*. Brisbane, Australia, 331–340.

[7] Peter Carr, Yaser Sheikh, and Iain Matthews. 2012. Point-less Calibration: Camera Parameters from Gradient-based Alignment to Edge Images. In *Proc. of IEEE Workshop on the Applications of Computer Vision (WACV '12)*. Breckenridge, CO, 377–384.

[8] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. 2014. Shape-Preserving Half-Projective Warps for Image Stitching. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. Columbus, OH, 3254–3261.

[9] Chun-Ming Chang, Cheng-Hsin Hsu, Chih-Fan Hsu, and Kuan-Ta Chen. 2016. Performance Measurements of Virtual Reality Systems: Quantifying the Timing and Positioning Accuracy. In *Proc. of the ACM Multimedia Conference (MM '16)*. Amsterdam, The Netherlands, 655–659.

[10] Tarek El-Ganainy and Mohamed Hefeeda. 2016. Streaming virtual reality content. *CoRR*, abs/1612.08350 (2016).

[11] Facebook. 2017. Surround 360. (2017). https://facebook360.fb.com/facebook-surround-360/.

[12] Christoph Fehn, Christian Weissig, Ingo Feldmann, Markus Muller, Peter Eisert, Peter Kauff, and Hans BloB. 2006. Creation of High-Resolution Video Panoramas of Sport Events. In *Proc. of the International Symposium on Multimedia (ISM'06)*. San Diego, CA, 291–298.

[13] FIFA Women's World Cup. 2015. FIFA TV broadcast production plan for FIFA Women's World Cup Canada 2015. (2015). http://resources.fifa.com/mm/document/tournament/competition/02/59/16/21/tvproductionbackgroundpaper_en_neutral.pdf.

[14] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

[15] Fujii Laboratory at Nagoya University. 2017. Nagoya University Multi-view Sequences Download List. (2017). http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/.

[16] Junhong Gao, Seon Joo Kim, and Michael S Brown. 2011. Constructing image panoramas using dual-homography warping. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. Colorado Springs, CO, 49–56.

[17] Marcel Germann, Tiberiu Popa, Richard Keiser, Remo Ziegler, and Markus Gross. 2012. Novel-View Synthesis of Outdoor Sport Events Using an Adaptive View-Dependent Geometry. *Computer Graphics Forum* 31, 2 (2012), 325–333.

[18] Bernard Ghanem, Tianzhu Zhang, and Narendra Ahuja. 2012. Robust video registration applied to field-sports video analysis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*. Kyoto, Japan, 1473–1476.

[19] GoPro. 2017. GoPro Omni. (2017). http://shop.gopro.com/virtualreality/.

[20] Ahmed Hamza and Mohamed Hefeeda. 2016. Adaptive Streaming of Interactive Free Viewpoint Videos to Heterogeneous Clients. In *Proc. of the ACM Multimedia Systems Conference (MMSys '16)*. Klagenfurt, Austria, 10.

[21] Kunihiko Hayashi and Hideo Saito. 2006. Synthesizing Free-Viewpoint Images from Multiple View Videos in Soccer Stadium. In *Proc. of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*. Sydney, Australia, 220–225.

[22] Michal Jancosek and Tomás Pajdla. 2011. Multi-view Reconstruction Preserving Weakly-supported Surfaces. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. Colorado Springs, CO, 3121–3128.

[23] Jaunt. 2017. Jaunt VR. (2017). https://www.jauntvr.com/.

[24] Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. 2011. Smoothly varying affine stitching. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. Colorado Springs, CO, 345–352.

[25] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[26] Cornelius Malerczyk. 2007. 3D-Reconstruction of Soccer Scenes. In *Proc. of the 3DTV Conference*. Kos Island, Greece, 1–4.

[27] Jörg Müller, Tobias Langlotz, and Holger Regenbrecht. 2016. PanoVC: Pervasive telepresence using mobile phones. In *Proc. of the IEEE Pervasive Computing and Communications (PerCom'16)*. Sydney, Australia, 1–10.

[28] Oculus. 2017. NextVR. (2017). http://www.nextvr.com/.

[29] Federico Perazzi, Alexander Sorkine-Hornung, Henning Zimmer, Peter Kaufmann, Oliver Wang, S Watson, and M Gross. 2015. Panoramic Video from Unstructured Camera Arrays. *Computer Graphics Forum* 34, 2 (2015), 57–68.

[30] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. *ACM Transactions on Graphics (TOG'03)* 22, 3 (2003), 313–318.

[31] Anne-Flore Nicole Marie Perrin, He Xu, Eleni Kroupi, Martin Řeřábek, and Tourajd Ebrahimi. 2015. Multimodal Dataset for Assessment of Quality of Experience in Immersive Multimedia. In *Proc. of the ACM Multimedia Conference (MM '15)*. Brisbane, Australia, 1007–1010.

[32] Svein Arne Pettersen, Dag Johansen, Håvard Johansen, Vegard Berg-Johansen, Vamsidhar Reddy Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Kvale Stensland, and Pål Halvorsen. 2014. Soccer Video and Player Position Dataset. In *Proc. of the ACM Multimedia Systems Conference (MMSys'14)*. Singapore, 18–23.

[33] Samsung. 2017. Project Beyondi. (2017). http://thinktankteam.info/beyond/.

[34] Sports TV Jobs. 2017. Live Sports Camera Positions and Responsibilities. (2017). http://www.infographicsarchive.com/sport/live-sports-camera-positions-and-responsibilities/.

[35] Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus A. Magnor. 2015. An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays. In *Proc. of the ACM Multimedia Conference (MM'15)*. Brisbane, Australia, 15–24.

[36] Håkon Kvale Stensland, Vamsidhar Reddy Gaddam, Marius Tennøe, Espen Helgedagsrud, Mikkel Næss, Henrik Kjus Alstad, Carsten Griwodz, Pål Halvorsen, and Dag Johansen. 2014. Processing Panorama Video in Real-time. *International Journal of Semantic Computing* 8, 2 (2014), 209–227.

[37] Ryo Suenaga, Kazuyoshi Suzuki, Tomoyuki Tezuka, Mehrdad Panahpour Tehrani, Keita Takahashi, and Toshiaki Fujii. 2015. A practical implementation of free viewpoint video system for soccer games. In *Proc. of the SPIE Three-Dimensional Image Processing, Measurement, and Applications*. San Francisco, CA, 93930G1–93930G8.

[38] Masayuki Tanaka, Ryo Kamio, and Masatoshi Okutomi. 2012. Seamless image cloning by a closed form solution of a modified poisson problem. In *Proc. of the ACM SIGGRAPH Asia Posters (SA'12)*. Singapore, 15.

[39] Daniel Wagner, Alessandro Mulloni, Tobias Langlotz, and Dieter Schmalstieg. 2010. Real-time panoramic mapping and tracking on mobile phones. In *Proc. of the IEEE Virtual Reality Conference (VR'10)*. Waltham, MA, 211–218.

[40] Changchang Wu. 2011. VisualSFM: A Visual Structure from Motion System. http://ccwu.me/vsfm/. (2011).

[41] Julio Zaragoza, Tat-Jun Chin, Quoc-Huy Tran, Michael S Brown, and David Suter. 2014. As-Projective-As-Possible Image Stitching with Moving DLT. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'14)* 36, 7 (2014), 1285–1298.

[42] Alireza Zare, Alireza Aminlou, Miska M. Hannuksela, and Moncef Gabbouj. 2016. HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications. In *Proc. of the ACM Multimedia Conference (MM'16)*. Amsterdam, The Netherlands, 601–605.

[43] Guofeng Zhang, Yi He, Weifeng Chen, Jiaya Jia, and Hujun Bao. 2016. Multi-Viewpoint Panorama Construction With Wide-Baseline Images. *IEEE Transactions on Image Processing* 25, 7 (2016), 3099–3111.