Recycling a Landmark Dataset for Real-time Facial Capture and Animation with Low Cost HMD Integrated Cameras

Caio José Dos Santos Brito
cjsb@cin.ufpe.br
Disney Research and U. Federal de Pernambuco

Kenny Mitchell
k.mitchell2@napier.ac.uk
Disney Research and Edinburgh Napier University

ABSTRACT

Preparing datasets for use in the training of real-time face tracking algorithms for HMDs is costly. Manually annotated facial landmarks are accessible for regular photography datasets, but introspectively mounted cameras for VR face tracking have incompatible requirements with these existing datasets. Such requirements include operating ergonomically at close range with wide angle lenses, low-latency short exposures, and near infrared sensors. In order to train a suitable face solver without the costs of producing new training data, we automatically repurpose an existing landmark dataset to these specialist HMD camera intrinsics with a radial warp reprojection. Our method separates training into local regions of the source photos, *i.e.*, mouth and eyes for more accurate local correspondence to the mounted camera locations underneath and inside the fully functioning HMD. We combine per-camera solved landmarks to yield a live animated avatar driven from the user's face expressions. Critical robustness is achieved with measures for mouth region segmentation, blink detection and pupil tracking. We quantify results against the unprocessed training dataset and provide empirical comparisons with commercial face trackers.

CCS CONCEPTS

• **Computing methodologies → Motion capture**; **VR**.

KEYWORDS

real-time, facial capture, virtual reality, HMD, data preparation

ACM Reference Format:

Caio José Dos Santos Brito and Kenny Mitchell. 2019. Recycling a Landmark Dataset for Real-time Facial Capture and Animation with Low Cost HMD Integrated Cameras. In *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI '19), November 14–16, 2019, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3359997.3365690

1 INTRODUCTION

Head mounted displays (HMDs) are used broadly in many applications [IDC 2017], such as animation [Olszewski et al. 2016], content creation [Vogel et al. 2018], medical applications [Egger et al. 2017], serious games [Gamito et al. 2017], object interaction [Figueiredo et al. 2018], and education [Dinis et al. 2017]. Despite the high popularity, for collaborative applications like games and video conferencing, which is necessary to visualize the face expressions of the user, it can be difficult to map the user's expression to the digitally animated avatar. Recent works including Suzuki *et al.* [2017] and Olszewski *et al.* [2016], tackle this issue by mounting sensors inside the HMD to capture motions from parts of the user's face. These works typically use machine learning approaches to estimate the face pose from sensor data, and require many captures of different users wearing the HMD to train the algorithm, demanding great manual effort to acquire each dataset.

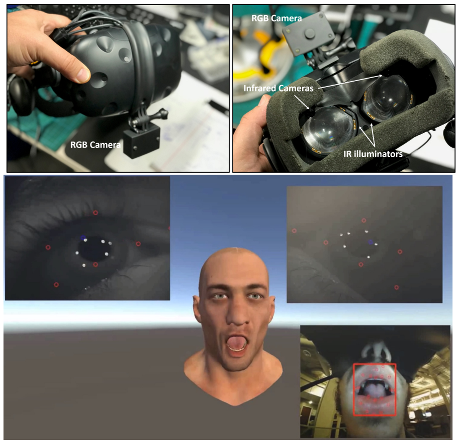

Figure 1: A virtual reality head mounted display (HMD) customised with specialized integrated cameras for real-time face tracking to drive an animated avatar (middle). The wide field of view color camera is mounted in close proximity underneath (RGB captured frame is shown in bottom right). In the top right, another view shows the binocular eye displays with integrated infrared cameras and illuminators. Eye camera captured images are shown beside the resulting avatar.

There exists several datasets of landmark labeled casual photography available online, such as the MTFL dataset [Zhang et al. 2014], the i-bug dataset [Sagonas et al. 2016, 2013a,b], and the Helen dataset [Le et al. 2012] that provide a large number of images with large varition of head poses and facial expressions, but they are not designed to be used in the HMD application because the images don't match the characteristics of the HMD mounted cameras. With each hand labeled dataset one can train a method to predict the landmark locations from a camera image of a face in real-time, such as Dlib's real-time face predictor [King 2009] [Kazemi and Sullivan 2014]. Our novel training dataset preparations are validated upon this facial landmark regression method for use in an HMD. We believe our method is the first to use such landmark based HMD face

VRCAI '19, November 14–16, 2019, Brisbane, QLD, Australia

ACM ISBN 978-1-4503-7002-8/19/11…$15.00
https://doi.org/10.1145/3359997.3365690

tracking. It is extendable to other face tracking algorithms such as Olszewski *et al.* [2016], as we apply warping on source face images prior to training, and apply further refinements in post, without altering the function of the core tracking solver.

With a *Pauschian* spirit [Pausch 1991], the main contribution of this paper lies in low cost recycling of an existing and accessible manually labeled image dataset built with regular photography to perform reasonably good landmark tracking in HMDs for avatar animation. In recycling an existing dataset we partition and adapt the data to a new target usage, essentially as a re-purposing of the valuable data to enable a new application. With an RGB camera viewing the lower face only and IR cameras viewing each eye, our contributions for recovering a live facial motion driven avatar within a fully functional virtual reality HMD are as follows,

- An automated warp reprojection to distort source training photos' camera characteristics to match the introspective HMD cameras' properties for each local facial region.

Whilst critical and key, our localized spherical training warp scheme was insufficient alone and we found the further contributions necessary to improve the overall system performance.

- Segmentation method based on histogram, HSV image and chromatism to calculate tight mouth bounding box reducing peripheral non mouth background/motion and in an area of reduced distortion more central to the near placement mouth camera lens.
- HSV channel segmentation for pupil position and blink detection.
- Dominant luminance channel optimization through direct (green) component use, rather than indirect color space conversion.
- Averaging and Kalman filtering, smoothing blendshape weights and decreasing temporal jitter present in landmark based solvers.
- Landmark based motion mapping to retarget calibrated neutral pose configurations to explicit expression pose activations.

The next section discusses the related work of image re-targeting and avatar face animation within HMDs, section 3 describes our camera setup used to capture the HMD user's eyes and lower face, and the method developed to warp the source dataset images and the mouth bounding box calculation. Section 4.2 describes the application used to validate the training data recycling method. Results and comparisons with non-warped dataset are detailed in section 5, and, finally, section 6 contains discussion and future work.

2 RELATED WORK

Our goal is to repurpose an image dataset of landmark labeled casual photography to match the distortion of distinct camera lenses attached to a HMD, and in turn to animate a 3D character using these landmarks tracked on the HMD user's face. The landmarks are predicted from live camera frames using the warped images as a recycled training dataset for the estimator. In the following, we review related works on this topic.

2.1 Image Retargeting

Image retargeting is the technique of resizing an image for a specific display or a different application without losing the content and structure of the region of interest (ROI). It is mainly used for adjusting video content to small displays and stereo video [Kiess et al. 2018]. Several works retarget images to use in customized applications. For instance, the work of Liu and Gleicher [2005] proposed an image retargeting method based on fisheye-view warping to adapt large images into small displays. The method automatically identifies a single region of interest (ROI) to emphasize, and it uses a non-linear image warping function to de-emphasize less important aspects of the image. To achieve better visibility on the foreground of omnidirectional images with a 360°field of view, Yu *et al.* [2018] proposed a to retarget the omnidirectional image into a 3D spherical image and assign the spherical nodes, which contains pixel values, using point correspondence between spherical polygons. Different from our approach which applies a radial based deformation, this method uses a point correspondence based spherical warp and does not aim to adapt changes in the device used. Further, the related works aim to minimize the distortion after retargeting the image, our work intentionally distorts the source cameras' images to match the deformation of the camera lenses attached to the HMD.

2.2 Avatar Face Animation within HMDs

Recently, Zollhöfer *et al.* [2018] presented an extensive state-of-art report on monocular 3D face reconstruction, tracking and applications. The report covers topics like facial capture with different input methods, optimizations, blendshape models, face reconstruction, and several applications. Our work doesn't intend to improve any individual state-of-art work on face tracking, but to demonstrate how generic datasets can be repurposed into different applications using any current or new tracking methods.

In this section, we focus on works that control a 3D character using the tracked expression of a person wearing a HMD. The first work on that research theme was proposed by Li *et al.* [2015] which combines strain signals a head-mounted RGBD camera image to track the mouth image. The system has a real time performance but requires a training calibration for each user, is not able to get the pupil location of the user and it struggles with blink detection. Furthermore, the results can present instability due to the pressure variation of the user's head and users suffer relatively poor ergonomics of the strain gauge assemblies.

Olszewski *et al.* [2016] presented a convolutional neural network (CNN) which regresses the lower face image of a person wearing a HMD and an IR image of the eyes to blendshapes weights that control the 3D character. The training dataset was created with a set of HMD users videos aligned with the face animation poses using audio-based alignment. The system can be used for real-time application, and it has a high fidelity result compared to other works, but it may not be robust to users that have a significant appearance difference compared to the training set. The training dataset used in the work is not available for the purpose of scientific comparison without reproduction of an equivalent dataset, which leads us to adopt the landmark regression approach [Suzuki et al. 2017] for validation of our dataset repurposing method.

Instead of using a mouth facing camera attached to the HMD, Song *et al.* [2018] presented a CNN based solution for real-time 3D face-eye performance with the camera looking at the whole face in non-head mounted configuration. A specific dataset was generated with labeled HMD face images and infrared eye images from multiple subjects. The results of this work are robust, but the method does not consider the eyelid movement, nor detects blinks. It relies on an externally aligned mouth facing cameras and fiducial marker detection to register the position limiting the range of motion of users. Finally, the system may not be robust to different lighting and different appearances from the statically captured HMD and specific manually created source training dataset.

With a similar setup to their own previous work, Zhao *et al.* [2019] proposed a framework to synthesize face images without the occlusions of the HMD. The framework is composed of four modules: 3D head reconstruction, face alignment and tracking, face synthesis, and eye synthesis, which can reconstruct the user’s face in, approximately, 500 milliseconds. The system requires to have a dataset with various prior poses of the user’s head to reconstruct the face, but as the camera setup is not integrated with fully functioning a head mounted display, it does not deal with head pose changes or user freedom of motion. Thies *et al.* [2018] introduced FaceVR, an image-based method which allows teleconferencing in VR based on self-reenactment. The camera setup has two IR cameras inside the HMD to capture the eye movements and an external RBG-D camera capturing the user‘s face. The system achieves real time performance, and it can reenact blinking and eye gaze but it is a person-specific system which (albeit quickly) requires to gather every users’ images. Related is the work of Rekimoto *et al.* [2018], but also requires person-specific scans. Suzuki *et al.* [2017] proposed a mapping using retro-reflective photoelectric sensor being able to estimate the character expression using five basic facial expressions (Neutral, Happy, Angry, Surprised, and Sad) that was used to train the neural network and had an overall accuracy of 88% in recognizing the facial expressions, but it requires a calibration for each user and has limited quality mouth animation due to the low number of sensors.

Finally, Lombardi *et al.* [2018] developed a data-driven rendering pipeline using a deep appearance model for rendering human faces of users wearing an HMD. The results are high quality, and the system runs in real time, but it necessitates a large number of per user’s face images, which are captured by a 40 camera setup capable of synchronous image ingest at 30 frames per second with 5120 x 3840 resolution, leading to an expensive solution. Despite the high-quality results, the system is only able to track and render users who are in the dataset, and it fails for other people. Wei *et al.* [2019] most recently develop this line of work with both *training* and *tracking* HMD camera configurations, resulting in similar high quality with lower costs, but still per user only. The training dataset used in the work is not available for the purpose of scientific comparison without reproduction of an equivalent dataset. We show our repurposing method employs a regression solver trained on an existing and freely available large dataset of general face photography, which performs reasonably for everyone immediately.

Figure 2: Samples from 15k hand labeled i-bug dataset[1].

3 METHOD

To achieve facial feature tracking with ergonomically mounted introspective cameras within a full functional VR HMD (detailed in section 4.1, we developed the following process: 1) source dataset warping to target camera intrinsics, 2) training of the shape detector using the new dataset for sub-regions, 3) additional mouth and eye detection refinements.

3.1 Dataset Warping

3.1.1 Facial Landmark Dataset. Our method recycles an existing dataset of $15k$ landmark labeled casual photography images of people’s faces with arbitrary poses from the variety of camera lenses to the target camera lenses of the HMD VR device. The source dataset is a set of $15k$ labeled casual photography of people faces from the i-bug dataset [1] [Sagonas et al. 2016, 2013a,b] with a large variation of head poses, skin color, mouth shapes, eye shapes and facial proportions which do not exactly match the distortion of the cameras mounted on and within the HMD. The images are in jpg format with various resolution, and the labeled landmarks are stored in a json file, which contains the 68 landmarks (17 for the jaw, 6 for each eye, 5 for each eyebrow, 9 for the nose, and 20 for the mouth). Samples of the source dataset can be seen in Figure 2.

3.1.2 Spherical Warping. To obtain a dataset with a closer target camera distortion, the mouth and eyes regions were cropped using the bounding box created by the position of a selective subset of local landmarks and scaled by an additional factor (10%) to match the corresponding proportions of the HMD camera images’ facial coverage area. The regions were cropped to get similar images as the ones captures by the HMD cameras: two eye images and the mouth image. Then, spherical warping is performed to create a final image with a similar target camera lens distortion.

We modeled lens distortion in radial terms per Equation 1 without accounting for tangential terms [MathWorks 2019].

$$\begin{aligned} P'_x &= P_x(1 + k_1 * r^2 + k_2 r^4 + k_3 * r^6) \\ P'_y &= P_y(1 + k_1 * r^2 + k_2 r^4 + k_3 * r^6) \end{aligned} \tag{1}$$

where P is the normalized pixel coordinates, P' is the radial distorted pixel coordinate, k_1, k_2, and k_3 are the radial distortion coefficients of the target camera lens intrinsics, and r is equal to $\sqrt{x^2 + y^2}$. The lens intrinsics are found through a regular chessboard reference pattern camera calibration process [Zhang 2000]. The radial distortion calculation is applied to the result of the spherical warp. Our spherical warp maps the source camera lens low distortion to the HMD cameras’ wide angle high distortion [Szeliski 2007].

$$\begin{aligned} (x_{dist}, y_{dist}) &= (s\theta, s\phi) + (x_c, y_c) \\ \theta &= \tan^{-1}(\frac{x}{f}) \\ \phi &= \tan^{-1}(\frac{y}{\sqrt{x^2 + f^2}}) \end{aligned} \tag{2}$$

where (x, y) are the spherical coordinates of an image pixel, (x_{dist}, y_{dist}) are the warped pixel coordinate, (x_c, y_c) are the warping center, s is the final image size, and f is the focal length. Given insignificantly low distortion of the cropped source photos in the eye and mouth regions, the distorted pixel can simply be calculated as in Equation 3.1.2 without a prior undistortion step. However, as the

[1] Licensed from Facesoft ltd, Annotated Facial Images, 2018.

eye cameras are mounted at an angle for best field of view coverage, the source image is rotated to match this 45°camera orientation, prior to warping (as shown in the eye reference image Figure 3).

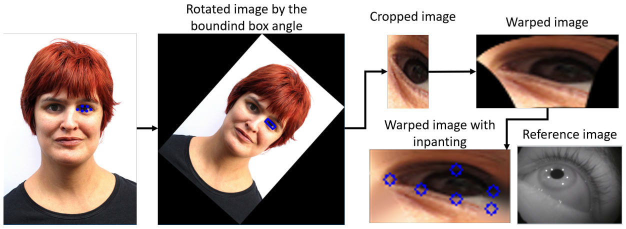
Rotated image by the boundind box angle
Cropped image
Warped image
Warped image with inpanting
Reference image

Figure 3: Eye spherical warp reprojection scheme.

Our warp is performed with center equal to the landmarks average position and with an Euler angle rotation of 0.8 radians (approx. 45°) on the x-axis only for the eye images to match the rounded pupil's shape (see Figure 3 top right warped image). Finally, an image inpainting process by Telea [2004] is performed to fill empty regions left due to the warp.

3.2 Shape Predictor Training

The target warped dataset was used to train Dlib's shape predictor [King 2009] which is an implementation of the ensemble of regression trees (ERT) method [Kazemi and Sullivan 2014]. For the purpose of comparability, training is per baseline parameters from Dlib: with the learning rate of 0.1 to avoid overfitting and suitable for our dataset size, tree depth of 4 for accuracy of the model, cascade depth of 10 and the number of trees per cascade equal to 500 yielding lower error through employing the cascade of regressors per [Kazemi and Sullivan 2014]. This approach has excellent real time performance and good quality predictions, but instead of using a single model for the whole face, two models with our focused warped images were trained, one for the mouth and one for each eye (with mirroring).

The dataset has 15k landmark labeled casual photography images of people's faces with arbitrary from different camera lenses varying the head pose, eyes opening, mouth shape, skin color. The mouth predictor was trained with 20 landmarks (12 for the outer lips and 8 for the inner lips), and the eye predictor was trained with 6 landmarks. Only a single model was trained for both eyes since we have a large number of eye samples with different shapes and photography styles. Each subject's right eye was used in training, and our run-time flipped the left eye camera image.

Also, the training was not done using the whole image, instead, the landmarks were used to calculate the bounding region of interest. The bounding box area is scaled by a 10% factor to match the real image proportion and it is used to crop the images to a new set with only the mouth and eye images. The bounding box is also used as a requirement for the landmarks prediction, the whole image is used as a bounding box for the eye predictor, and the mouth bounding box calculation follows in the next section.

3.3 Mouth Region Estimation

We localise landmark prediction processing using a mouth bounding box estimation method combining an adaptive threshold algorithm by Panning *et al.* [2009] (which was also recently applied to mobile VR by Teng and Yang [2016]) with a chromatism based lips segmentation [Ji et al. 2018].

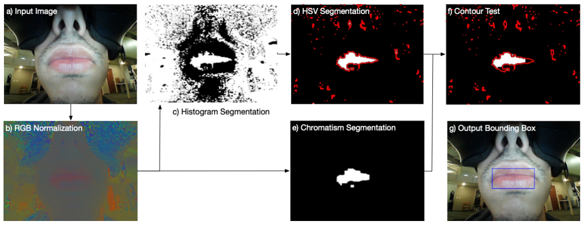

Figure 4: Mouth bounds estimation. a) input, b) color normalization, c) histogram segmentation, d) our HSV segmentation, e) chromatism segmentation, f) our contour, g) result.

The lower face image is first color normalized (Figure 4b) and used as input for three types of segmentation. Given an image with pixels p and a histogram with bin value $h(i)$, a pixel is classified as a skin pixel if is follows adaptive threshold range conditions to avoid false positives [Panning et al. 2009] (Figure 4c). Then, to find the mouth pixels, the image is converted to HSV color space, and it is filtered by the hue channel between 0 and 120 (for the range of our test subjects wearing no unusual lipstick), and it is subtracted from the histogram based segmented image, as visualized in Figure 4d. Then, the contours are extracted from the segmented image to calculate the mouth area using a convex hull approximation.

These filter steps are able to segment the mouth region, but some noise might occur in the image which may lead to an incorrect mouth region segmentation with multiples contours found (Figure 4d). To improve the result and choose the correct contour, the RGB color space chromatism lips segmentation is performed (Figure 4e). The chromatism value s is calculated by Equation 3 and has a value greater than 0 for pixels in the lips region [Ji et al. 2018].

$$s = 2\,tan^{-1}(\frac{R-G}{R})/\pi \tag{3}$$

The chromatism segmentation is achieved after performing dilation and erosion operations to remove small noise artifacts caused by the low latency short exposure camera, but this method alone is only able to segment a small portion of the mouth. Therefore, the segmentations are combined by testing which contour center is inside the chromatism lips segmentation (Figure 4f). Finally, the mouth bounding box is calculated using the maximum and minimum contours points (Figure 4g).

3.4 Blink Detection and Pupil Estimation

The eye shape predictor is able to estimate the eyes landmarks, but it has two main issues: it is not reliable to estimate the eyes landmarks for closed eyes due to the majority of training photos having eyes open (Figure 5), and it is not able to stably estimate the pupil position.

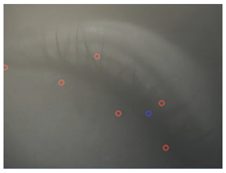

Figure 5: Incorrect landmark prediction for a closed eye.

To detect an eye blink, the reflection of the IR illuminators attached to the HMD lens into the eyes is measured. This is accomplished by converting the eye image to HSV color space and performing a thresholding operation, where more than 10 pixels must fall within the range of [220-255] of the V channel.

To address the pupil estimation issue, another thresholding operation is performed to segment the HSV eye image and find the pupil region that filters the V channel in a range between 0 and 10. Then, to calculate the center of the pupil, the moment is calculated based on the work of Williams [1990] for pixels inside the eye landmark bounding box, and the optimization proposed by Spieldenner *et al.* [2014], which minimizes inner loop detection comparisons, was used to achieve high computational performance (see Figure 6).

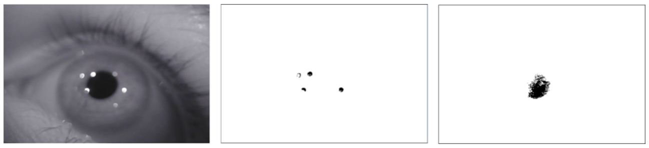

Figure 6: Input image (left), IR illuminators reflection segmentation (middle), and the pupil segmentation (right).

4 IMPLEMENTATION

4.1 Ergonomically Integrated Cameras in a Fully Functioning HMD

The experimental camera setup is composed of three cameras attached to the headset display (see Figure 1). The first camera is a RGB camera OmniVision OV9712 attached on the base with a 3D printed adjustable assembly, which can visualize the lower face. It has a 320 x 240 image resolution connected via USB, a 120° wide angle lens. The wide lens permits the camera to position directly under the display very close to the mouth, which provides a greater freedom of movement than sensors mounted further away protruding from the display [BinaryVR 2015] [Strassburger 2018]. The pair of 120Hz eye tracking infrared cameras provided by Pupil Labs [Kassner et al. 2014] include two rings with IR illuminators attached around the display lenses. These cameras also have 320 x 240 image resolution connected via USB. The processing hardware used to run the application had an Intel Xeon CPU E5-2630 v4, 16 GB memory, and a NVIDIA 4G Quadro K2200. Two different 3D face sample characters were used to test our application which is able to perform the artist designed expressions (Figure 7).

4.2 Application

For a demonstrative application of the repurposing technique, the mouth and eyes landmarks were used as input to animate the face of a 3D character which is composed of a set of blendshapes. The prototype application can generate six expressions on the character: open and close the mouth, move the mouth in the horizontal direction, create a smile and funnel shape, open and close the eyes, and look to the left and right.

The application was developed on top of the Faceware Tech Live sample [Tech 2018] on Unity 3D version 2018.2.4. However, instead of using the Faceware Live server to calculate the blendshapes weights which does not work for our application's cameras (see Figure 16), it uses our proposed method to predict the landmarks and calculate the blendshapes weight.

In applying the tracked landmarks to corresponding avatar motion we perform a rig calibrated motion mapping. First, the calibration distances between neutral pose face landmarks are stored to demarcate their configuration and correspondingly we set the blendshapes weight to form the matching neutral expression (this is typically a set of zero blendshape activations for a regular face rig). Then for each captured frame, new face landmark positions are estimated using our recycled feature prediction method, and the new landmark distances are calculated. Finally, the new blendshapes weights are activated by according to a linear interpolation between the calibration distances and the new landmarks distances.

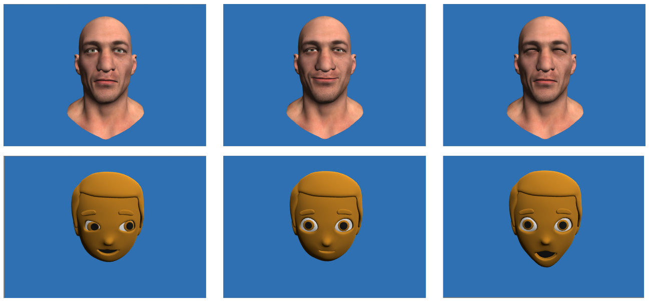

Figure 7: Expressions that our simple driven character demonstration can perform. Top: neutral (left), smile (middle) and eyes closed (right). Bottom: looking right (left), funnel shape (middle) and mouth open (right).

Given that the image sensor data ingest with Unity is in RGB color space, instead of performing a conversion to either HSV or YCbCr spaces to segment the pupil image, the thresholding operation is performed using the green channel of the image for more direct sensor data whilst yielding the same consistent result.

To improve the temporal behavior and decrease the presence of jitter on the mouth prediction results, a Kalman filter operation was done to smooth the bounding box estimation over time. The filter was used in the top left point and the bottom right point of the bounding box with process a noise covariance matrix equals to a diagonal matrix with value of 0.0001, a measurement noise covariance diagonal matrix with value of of 0.1, and posteriori error estimate covariance diagonal matrix with with value of of 0.1. Finally, an average of two consecutive values were used to smooth the blendshapes weight calculation.

5 RESULTS AND ANALYSIS

The repurposing method generates new mouth images with similar distortion to the HMD lower face image which shows prominent lips on the center of the image (Figure 8). The source dataset has different lips size and shapes that allow a target dataset with high variability that is used to train the Dlib's shape predictor.

The method is also able to reproduce the camera distortion of the infrared eye camera which captures the eyes with the angle close to 45° on the bounding box and with the curved shape for the top eyelid. The recycled eye images can be seen in Figure 8.

Using these recycled images as training images, the method predicts 20 mouth landmarks and 6 landmarks for each eye. The technique predicts the landmarks for different lips sizes and shapes but may generate symmetrical mouth shape results (Figure 9). Also, the eye landmarks predict the eye shape with the correct eyelid curvature, but it does not present a high accuracy which may be because of the fixed bounding box (Figure 13).

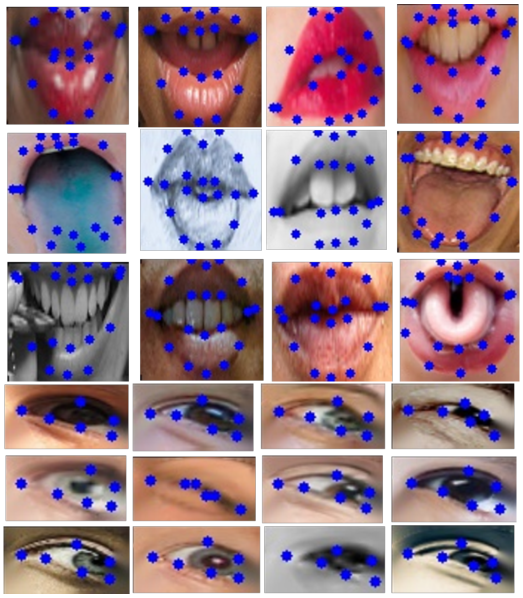

Figure 8: Samples of the extracted and warped mouth and eye images from the source training dataset.

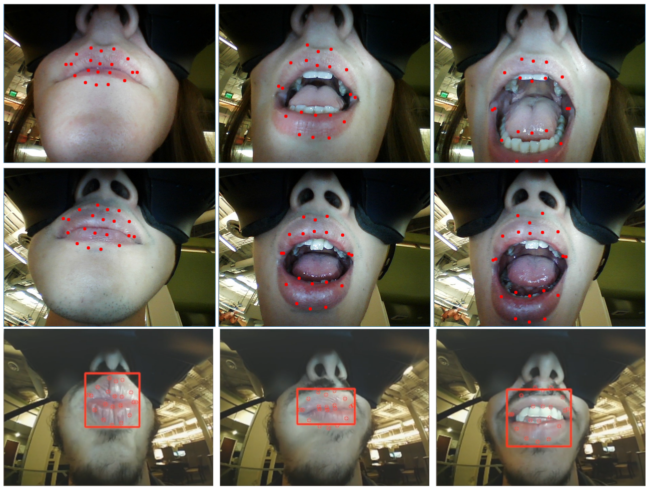

Figure 9: Result of the mouth region and lip landmark predictions for various different lips poses.

A time profile in ms was measured to identify which steps of the mouth landmarks prediction were most time consuming, the results can be seen in Table 1. The chromatism segmentation and the histogram were the most time consuming steps of the mouth prediction due the manipulation of two images in these stages. The histogram segmentation result is used to filter the HSV segmentation, and the HSV segmentation result is used as filter for the chromatism segmentation to avoid miscalculation of bounding box.

5.1 Prototype Application Evaluation

The application was tested with four different people with distinct skin color and lips size. In the test, the user spoke several phrases including some pangrams to reach the whole spectrum of the alphabet, for instance, "The five boxing wizards jump quickly", and the face blendshapes weight is calculated using the difference between the current landmarks and the neutral expression landmarks.

Table 1: Time profile for the mouth bounding box calculation and landmarks prediction.

Step	Time (ms)
RGB Normalization	4.29
Histogram Segmentation	8.91
HSV Segmentation	2.31
Chromatism Segmentation	15.51
Landmarks Prediction	1.98

The method animates avatar frames at a steady 30fps in our experimental setup and the results can be seen in Figure 10. We observed stable pupil prediction and blink detection, able to handle different eyelid openings Figure 11 (left) and different illuminations Figure 11 (middle), but it does not precisely provide the central pupil position Figure 11 (right). Also, the Kalman filter is not able to fully prevent the presence of jitter over the frames without introducing noticeable lag, which is also a common aspect of the employed landmark regression solver's behaviour on regular cameras, and our reprojection preparation performs no worse than this.

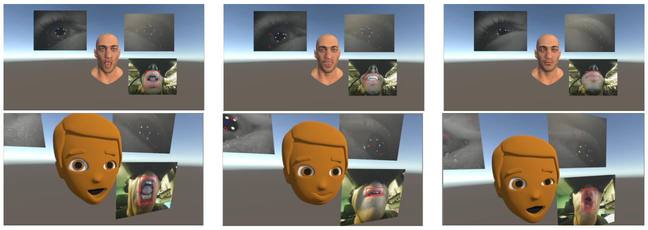

Figure 10: Victor and Emoji face characters driven by the face landmarks of the HMD user in real-time live.

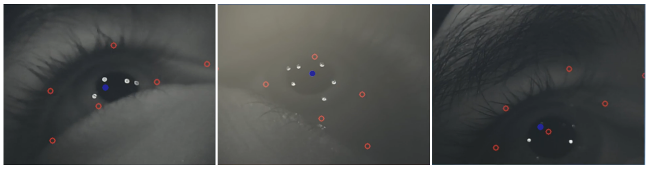

Figure 11: Pupil detection results (blue circle).

In comparison to previous solutions, our work does not rely on acquiring a new training dataset custom to HMD cameras or any dataset for calibration and it can be replicated using any other labeled casual photography dataset or training set for any face tracking algorithm. Our use of the regression solver does not present a high fidelity of output with jitter on the detection results and simple poses driven directly from the tracked landmarks. However, repurposing training data for use with alternative solvers [Zollhöfer et al. 2018] would follow our method, e.g. drawing from the coarse and refinement approaches of Ma and Deng [2019]. The approach taken to drive blendshapes weights is much cheaper than solving a landmark driven animation retargeting custom to each avatar rig, but this is beyond the scope of this article's focused contributions.

5.2 Comparison with non repurposed method

5.2.1 Ablation Test. First, the standard landmark predictor was tested without the proposed mouth bounding box calculation, and the whole image was used as the bounding region of interest for the Dlib's shape predictor. Without the correct bounding box calculation, the shape predictor calculates a completely wrong result (Figure 12 left). Our method detailed in subsection 3.3 calculates the landmark with better positioning covering the mouth aperture and lips shape (Figure 12 right).

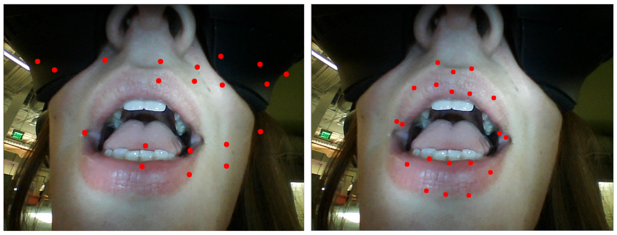

Figure 12: Ablation test comparison between the mouth prediction result using standard predictor with no bounding box calculation (left) and our method (right).

5.2.2 Quantitative Comparisons. Our method provided a more precise mouth landmark predictions in comparison to standard predictor being able to fit the landmarks position closer to the lips' true contours. Often our method performs better on the lower lip landmarks, but it reduces precision in a few cases (see Figure 13). Here the quantitative pixel distance error of our method versus hand labeled ground truth over the mouth camera resolution 320x240 is 31.34 versus 37.0 for prediction with the unprepared training dataset, an 18% improvement.

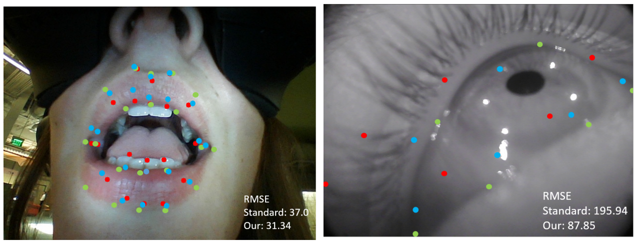

Figure 13: RMSE Comparison of the mouth and eye prediction results using our method and the standard predictor. Blue points are the landmarks using our method (mouth pixel distance RMSE is 31.34 and eye RMSE is 87.85), red points are the landmarks using the standard predictor (mouth RMSE 37.0 and eye RMSE 195.94), and the green points are the manually labeled ground truth landmarks.

The standard predictor achieved a wrong eye prediction, and whilst it is not perfect, our method finds the correct localization for the eye, being successful in adjusting the landmarks shape to the eyelids' curvature. Quantitatively, our method results with root mean square error of 87.85, while the standard predictor achieves an RMSE of 195.94 (Figure 13).

No quantitative comparison was performed with the HMD face tracking systems listed on the related works section because, due to our knowledge, there is no open source implementation of those methods and we do not have access to the setup used in the researches. However, a qualitative comparison with the state of the art works shows that our work does not present highly detailed results due to our choice of sparse landmark prediction method, but it is a much faster solution that estimates mouth landmarks, pupil movement, and blink detection for any face without requiring user-specific training data. Our solution can be applied to different labeled facial datasets, such as the MTFL dataset [Zhang et al. 2014], the i-bug dataset [Sagonas et al. 2016, 2013a,b], and the Helen dataset [Le et al. 2012] to create recycled training dataset for different face tracking algorithms.

5.3 Empirical comparisons with commercial applications

An important class of face tracking related to this work is mobile device tracking with small form factor cameras potentially suited to our HMD face tracking task. Strassburger [2018] performs real-time face tracking using an iPhoneX with RGB-D sensor. In Strassburger's demonstration, the device is head mounted but some distance away from the mouth and does not include eye tracking. Three commercial mobile apps that drive a face character using facial tracking were tested to analyze their performance and suitability for mouth tracking with a worn HMD.

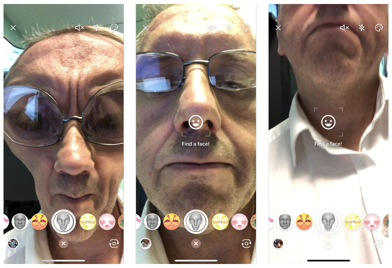

Figure 14: Facebook Messenger Emojii app where tracking succeeds at reasonably close range (left), but fails for a face distance lower than 15*cm* (middle), and it does not work for lower face images only (right).

The first is the Facebook Messenger Emoji app which uses depth based face tracking camera hardware, being able to predict the face shape for a minimum distance of 15*cm* between the user's face and the phone, and it is not able to predict the face landmarks of only the lower face region. The results presented a instability on the face movements because of the jitter, and can be visualized in Figure 14.

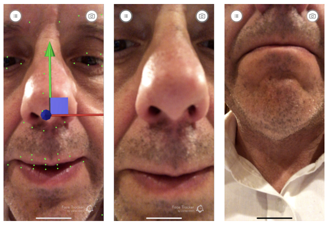

Figure 15: Image based Pinscreen app where their algorithm is able to predict the face landmarks for a near face (left), but fails for distances closer than 8*cm* (middle) and when only the mouth is visible.

Two versions of the Pinscreen app [Pinscreen 2018b] were tested: the Pinscreen App [Pinscreen 2018a] performs face tracking using the depth sensor of the iPhoneX and it has a similar performance to the Facebook Messenger Emoji app, being able to perform the face tracking within a 15*cm* distance and is not able to track the lower face image. The Pinscreen Face Tracker [Pinscreen 2018a] app is an image based tracking that has a smaller minimum distance of 8*cm* but it is still not able to

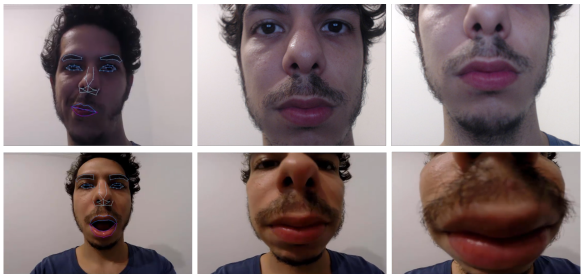

Figure 16: Faceware Live succeeds for full face (left), but fails when closer than $15cm$ (middle), or lower face only is visible (right). Top: C920 webcam. Bottom: GoPro.

track only the lower face region, as can be viewed in Figure 15. Both tracking systems presented jitter over time, making the face movements a little unstable.

The Faceware Live desktop application was also tested using a Logitech C920 HD webcam and a GoPro Hero 3+ Black Edition that has a fisheye lens with a similar distortion to the camera on the bottom of the HMD. The Faceware tracking was able to predict the landmarks position for both cameras when the user is at a minimum distance from the camera of $15cm$ for the webcam and $4cm$ for the GoPro, and the tracking fails when the image only shows the lower face part. The test result can be seen in Figure 16 and the movements of the face character were stable over time. Although this is not a direct quantitative comparison with our proposed method, this empirical comparison shows that lower face tracking completely fails on existing commercial solutions that were tested and our work can be used as basis to solve this issue with a variety of choices for the facial tracking solver algorithm using our training data recycling method.

6 DISCUSSION AND FUTURE WORK

This paper presents a radial warp based image retargeting to match casual photography labeled images to the lens distortion of cameras integrated into a low cost HMD. This data preparation method avoids the cost for manually creating a large training landmark dataset by recycling an existing one using our image retargeting.

The warped labeled images were used as the training dataset for the Dlib's [King 2009] shape predictor which was able to predict the face landmarks for different lips size and shapes. The 15k training images for both eye and mouth tracking provided a good deal of variation, but further variation could have been introduced via perturbation of warp parameters, generating left eye training images and so on. Nonetheless, our method was able to achieve a better result on the mouth and eyes prediction in comparison to the same predictor using the unprocessed source dataset. The proposed method can be used in many face tracking algorithms that rely on a labeled image training set, such as the work of Ren *et al.* [2017], and the work of Yu and Luo [Yu et al. 2016] which are based on CNNs. The method can be applied to other uses such as robotics [Courbon et al. 2007] and autonomous vehicles [Bertozzi et al. 2015] where specialised camera intrinsics are also often employed.

We warp the training dataset to match the live capture cameras' intrinsics, however, it may be an interesting direction to explore more sophisticated warps of both live image data in combination with existing training photography to meet an optimum salient facial feature space for ideal capture accuracy.

A tight mouth region calculation based on histogram, HSV channel and chromatism segmentation is proposed to more accurately localise the landmark prediction region. One limitation of the mouth region calculation exists for cases where the background or the HMD user skin tone is close to the lips color, the segmentation may not be able to correctly isolate the mouth region. As with any landmark regression, tracking may also fail when the mouth is partially occluded. One possibility to address these limitations is to use a different color space like YCbCr as in the the work of Shaik *et al.* [2015] and also in the work of Yadav and Nain [2015], another possibility is using neural network approaches such as in the work of Zaidan *et al.* [2014], but it is necessary to have a specific training set for the method which may lead to a expensive solution.

As the eye training dataset RGB images are in different sensor space compared to the IR eye cameras inside the HMD, the detector might have a greater performance if trained with images that match the noise and spectral characteristics of the captured images. To decrease the domain gap between training data and application images, generative adversarial networks (GAN) have been proposed which learn to map between synthetic and real images with Rad *et al.* [Rad et al. 2019], Mueller *et al.* [2018], and Zakharov *et al.* [Zakharov et al. 2018]. This approach could be used in our solution but it is necessary to have a different dataset to train the GAN network to learn the mapping between RBG and infrared images.

A facial animation application which used the face landmarks to animate the 3D face character was constructed to validate the repurposing method. The application was able to generate several face expressions in real time, such as open and close the mouth, move the mouth in the horizontal direction, create a smile and funnel shape, open and close the eyes, and look to the left and right, but, as the blendshapes calculations were done manually, the digital avatar does not have high quality expressions over time with the presence of jitter. Commercial apps and software that predict the face landmarks to animate a digital character were tested with different cameras. One such configuration in head mounted face tracking by Strassburger [2018] uses a method similar to commerical apps with an iPhoneX RGBD camera, but is not able to be mounted as closely as our method, therefore exhibits poor ergonomics. We have shown the commerical methods are not able to predict the face landmarks for the lower face image at close range and it was noticed the presence of jitter in these applications.

The face animation application can be improved by decreasing the jitter, which was not fully achieved with our use of a Kalman filter. The blendshapes weight may be calculated directly, as solved by a neural network such as in the work of Olszewski *et al.* [2016] or indeed Wei *et al.* [2019], but remains to be seen whether these methods apply generically with a huge training dataset to provide HMD face tracking for any user and target rig.

ACKNOWLEDGMENTS

We thank the reviewers for their careful feedback. Mark Mine of Walt Disney Studios provided much encouragement. Ben Grossmann and Jason Crosby of Magnopus helped fabricate and assemble the headset camera mounts used in our user testing experiments.

REFERENCES

M. Bertozzi, L. Castangia, S. Cattani, A. Prioletti, and P. Versari. 2015. 360Âř Detection and tracking algorithm of both pedestrian and vehicle using fisheye images. In *2015 IEEE Intelligent Vehicles Symposium (IV)*. 132–137. https://doi.org/10.1109/IVS.2015.7225675

BinaryVR. 2015. VR Headset Calibration Mode. https://www.youtube.com/watch?v=yr2fFeympKY.

Jonathan Courbon, Youcef Mezouar, Laurent Eckt, and Philippe Martinet. 2007. A generic fisheye camera model for robotic applications. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 1683–1688.

F. M. Dinis, A. S. Guimarães, B. R. Carvalho, and J. P. P. Martins. 2017. Development of virtual reality game-based interfaces for civil engineering education. In *2017 IEEE Global Engineering Education Conference (EDUCON)*. 1195–1202. https://doi.org/10.1109/EDUCON.2017.7943000

Jan Egger, Markus Gall, JÃijrgen Wallner, Pedro Boechat, Alexander Hann, Xing Li, Xiaojun Chen, and Dieter Schmalstieg. 2017. HTC Vive MeVisLab integration via OpenVR for medical applications. *PLOS ONE* 12, 3 (03 2017), 1–14. https://doi.org/10.1371/journal.pone.0173972

Lucas Figueiredo, Eduardo Rodrigues, João Teixeira, and Veronica Techrieb. 2018. A comparative evaluation of direct hand and wand interactions on consumer devices. *Computers & Graphics* 77 (2018), 108 – 121. https://doi.org/10.1016/j.cag.2018.10.006

Pedro Gamito, Jorge Oliveira, Carla Coelho, Diogo Morais, Paulo Lopes, José Pacheco, Rodrigo Brito, Fabio Soares, Nuno Santos, and Ana Filipa Barata. 2017. Cognitive training on stroke patients via virtual reality-based serious games. *Disability and Rehabilitation* 39, 4 (2017), 385–388. https://doi.org/10.3109/09638288.2014.934925 arXiv:https://doi.org/10.3109/09638288.2014.934925 PMID: 25739412.

IDC. 2017. Worldwide Shipments of AR/VR Headsets Maintain Solid Growth Trajectory in the Second Quarter, According to IDC. https://www.idc.com/getdoc.jsp?containerId=prUS43021317. [Online; accessed 25-November-2018].

Yingyu Ji, Wang Wang, Yang Lu, Jian Wei, and Yan Zhao. 2018. Eye and mouth state detection algorithm based on contour feature extraction. *Journal of Electronic Imaging* 27, 5 (2018), 1 – 8 – 8. https://doi.org/10.1117/1.JEI.27.5.051205

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

V. Kazemi and J. Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874. https://doi.org/10.1109/CVPR.2014.241

Johannes Kiess, Stephan Kopf, Benjamin Guthier, and Wolfgang Effelsberg. 2018. A Survey on Content-Aware Image and Video Retargeting. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 3, Article 76 (July 2018), 28 pages. https://doi.org/10.1145/3231598

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. 2012. Interactive Facial Feature Localization. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 679–692.

Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial Performance Sensing Head-mounted Display. *ACM Trans. Graph.* 34, 4, Article 47 (July 2015), 9 pages. https://doi.org/10.1145/2766939

Feng Liu and Michael Gleicher. 2005. Automatic Image Retargeting with Fisheye-view Warping. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology (UIST '05)*. ACM, New York, NY, USA, 153–162. https://doi.org/10.1145/1095034.1095061

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages. https://doi.org/10.1145/3197517.3201401

Luming Ma and Zhigang Deng. 2019. Real-time Hierarchical Facial Performance Capture. In *Symposium on Interactive 3D Graphics and Games (I3DâĂŹ19)*. Montreal, QC, Canada. ACM, New York, NY, USA, 10. https://doi.org/10.1145/3306131.3317016

MathWorks. 2019. *What Is Camera Calibration?* https://www.mathworks.com/help/vision/ug/camera-calibration.html

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11. https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity Facial and Speech Animation for VR HMDs. *ACM Trans. Graph.* 35, 6, Article 221 (Nov. 2016), 14 pages. https://doi.org/10.1145/2980179.2980252

Axel Panning, Robert Niese, Ayoub Al-Hamadi, and Bernd Michaelis. 2009. A new adaptive approach for histogram based mouth segmentation. *Proceedings of the World Academy of Science, Engineering and Technology* 56 (2009), 779–784.

Randy Pausch. 1991. Virtual Reality on Five Dollars a Day. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*. ACM, New York, NY, USA, 265–270. https://doi.org/10.1145/108844.108913

Pinscreen. 2018a. Pinscreen App: Instant 3D Avatars. https://itunes.apple.com/us/app/pinscreen-instant-3d-avatars/id1322184255. [Online; accessed 25-November-2018].

Pinscreen. 2018b. Pinscreen: Instant 3D Avatars. https://www.pinscreen.com. [Online; accessed 25-November-2018].

Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2019. Domain Transfer for 3D Pose Estimation from Color Images Without Manual Annotations. In *Computer Vision – ACCV 2018*, C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (Eds.). Springer International Publishing, Cham, 69–84.

Jun Rekimoto, Keishiro Uragaki, and Kenjiro Yamada. 2018. Behind-the-mask: A Face-through Head-mounted Display. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces (AVI '18)*. ACM, New York, NY, USA, Article 32, 5 pages. https://doi.org/10.1145/3206505.3206544

Z. Ren, S. Yang, F. Zou, F. Yang, C. Luan, and K. Li. 2017. A face tracking framework based on convolutional neural networks and Kalman filter. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. 410–413. https://doi.org/10.1109/ICSESS.2017.8342943

Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 Faces In-The-Wild Challenge: database and results. *Image and Vision Computing* 47 (2016), 3 – 18. https://doi.org/10.1016/j.imavis.2016.01.002 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 2013a. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In *2013 IEEE International Conference on Computer Vision Workshops*. 397–403. https://doi.org/10.1109/ICCVW.2013.59

C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 2013b. A Semi-automatic Methodology for Facial Landmark Annotation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 896–903. https://doi.org/10.1109/CVPRW.2013.132

Khamar Basha Shaik, P. Ganesan, V. Kalist, B.S. Sathish, and J. Merlin Mary Jenitha. 2015. Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space. *Procedia Computer Science* 57 (2015), 41 – 48. https://doi.org/10.1016/j.procs.2015.07.362 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

Guoxian Song, Jianfei Cai, Tat-Jen Cham, Jianmin Zheng, Juyong Zhang, and Henry Fuchs. 2018. Real-time 3D Face-Eye Performance Capture of a Person Wearing VR Headset. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, New York, NY, USA, 923–931. https://doi.org/10.1145/3240508.3240570

T. Spieldenner, K. Sons, M. Lancelle, F. ZÃijnd, and K. Mitchell. 2014. Web3D 2014 tutorial: An Ecosystem for Interactive Mixed-Reality Applications on the Web. http://presentations.web3d.org/2014/Web3D2014/Tutorials/EcosystemInteractiveMixedRealityApplications/web3d2014.htm. [Online; accessed 25-November-2018].

Cory. Strassburger. 2018. Democratising Mocap: Real-Time Full-Performance Motion Capture with an iPhone X, Xsens, IKINEMA, and Unreal Engine. *ACM SIGGRAPH: Real Time Live!* (2018). https://www.youtube.com/watch?v=lXZhgkNFGfM

K. Suzuki, F. Nakamura, J. Otsuka, K. Masai, Y. Itoh, Y. Sugiura, and M. Sugimoto. 2017. Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)*. 177–185. https://doi.org/10.1109/VR.2017.7892245

Richard Szeliski. 2007. Image Alignment and Stitching: A Tutorial. *Foundations and Trends® in Computer Graphics and Vision* 2, 1 (2007), 1–104. https://doi.org/10.1561/0600000009

Faceware Tech. 2018. *Faceware Live: Professional Realtime Animation Software*. http://facewaretech.com/products/software/realtime-live/

Alexandru Telea. 2004. An image inpainting technique based on the fast marching method. *Journal of graphics tools* 9, 1 (2004), 23–34.

Teng Teng and Xubo Yang. 2016. Facial Expressions Recognition Based on Convolutional Neural Networks for Mobile Virtual Reality. In *Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry - Volume 1 (VRCAI '16)*. ACM, New York, NY, USA, 475–478. https://doi.org/10.1145/3013971.3014025

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2018. FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. *ACM Trans. Graph.* 37, 2, Article 25 (June 2018), 15 pages. https://doi.org/10.1145/3182644

Daniel Vogel, Paul Lubos, and Frank Steinicke. 2018. AnimationVR - Interactive Controller-based Animating in Virtual Reality. http://basilic.informatik.uni-hamburg.de/Publications/2018/VLS18

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (July 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Lance Williams. 1990. Performance-driven Facial Animation. *SIGGRAPH Comput. Graph.* 24, 4 (Sept. 1990), 235–242. https://doi.org/10.1145/97880.97906

S. Yadav and N. Nain. 2015. Fast Face Detection Based on Skin Segmentation and Facial Features. In *2015 11th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. 663–668. https://doi.org/10.1109/SITIS.2015.91

A. Yu, K. Hara, K. Inoue, and K. Urahama. 2018. Foreground Enlargement of Omnidirectional Images by Spherical Trigonometry. In *2018 24th International Conference on Pattern Recognition (ICPR)*. 2630–2635. https://doi.org/10.1109/ICPR.2018.8545697

H. Yu, Z. Luo, and Y. Tang. 2016. Transfer Learning for Face Identification with Deep Face Model. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. 13–18. https://doi.org/10.1109/CCBD.2016.014

A.A. Zaidan, N.N. Ahmad, H. Abdul Karim, M. Larbani, B.B. Zaidan, and A. Sali. 2014. Image skin segmentation based on multi-agent learning Bayesian and neural network. *Engineering Applications of Artificial Intelligence* 32 (2014), 136 – 150. https://doi.org/10.1016/j.engappai.2014.03.002

S. Zakharov, B. Planche, Z. Wu, A. Hutter, H. Kosch, and S. Ilic. 2018. Keep it Unreal: Bridging the Realism Gap for 2.5D Recognition with Geometry Priors Only. In *2018 International Conference on 3D Vision (3DV)*. 1–11. https://doi.org/10.1109/3DV.2018.00012

Zhengyou Zhang. 2000. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (December 2000), 1330–1334. MSR-TR-98-71, Updated March 25, 1999.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial Landmark Detection by Deep Multi-task Learning. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 94–108.

Yajie Zhao, Qingguo Xu, Weikai Chen, Jun Xing, Chao Du, Xinyu Huang, and Ruigang Yang. 2019. Mask-off: Synthesizing Face Images in the Presence of Head-mounted Displays. In *2019 IEEE Virtual Reality (VR)*. IEEE.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 523–550. https://doi.org/10.1111/cgf.13386