# Real-time expression-sensitive HMD face reconstruction

Xavier P. Burgos-Artizzu *, Julien Fleureau, Olivier Dumas, Thierry Tapie, François LeClerc, Nicolas Mollet
Technicolor, Cesson-Sévigné, France

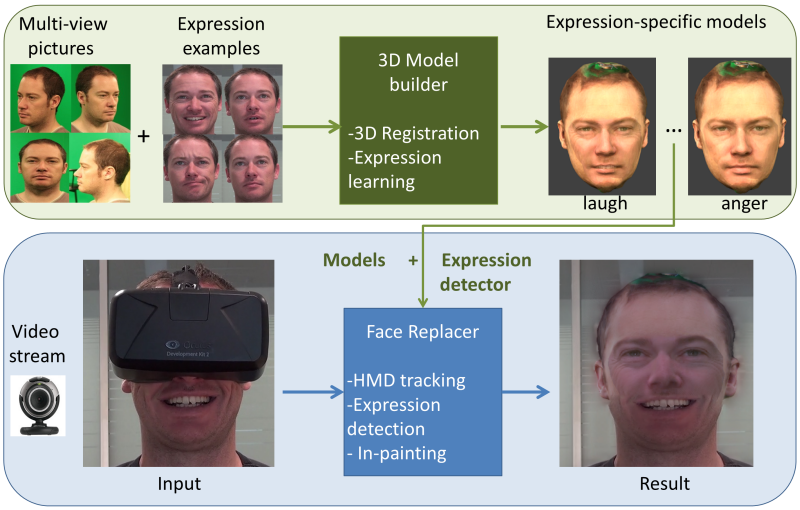



**Figure 1:** *Our method is able to reconstruct the face of a person wearing a Head-Mounted Display (HMD) in real-time. It does so while trying to respect the user facial expression, without the need for any other sensor besides a standard color video camera.*

## Abstract

One of the main issues of current Head-Mounted Displays (HMD) is that they hide completely the wearer's face. This can be an issue in social experiences where two or more users want to share the 3D immersive experience. We propose a novel method to recover the face of the user in real-time. First, we learn the user appearance off-line by building a 3D textured model of his head from a series of pictures. Then, by calibrating the camera and tracking the HMD's position in real-time we reproject the model on top of the video frames mimicking exactly the user's head pose. Finally, we remove the HMD and replace the occluded part of the face in a seamingless manner by performing image in-painting with the background. We further propose an extension to detect facial expressions on the visible part of the face and use it to change the upper face model accordingly. We show the promise of our method via some qualitative results on a variety of users.

**CR Categories:** 3.20 [Image & Video]: Texture Synthesis and Inpainting— [3.12]: Image & Video—Image processing; 3.1 [Image & Video]: 2D Morphing and Warping—;

**Keywords:** Face reconstruction, Head-mounted display

*e-mail:xavier.burgos@technicolor.com

## 1 Introduction

Not surprisingly, mass-market Head-Mounted Displays (HMD) such as Oculus Rift [OCULUS VR ], are becoming increasingly popular. After major design improvements during the last years, they are now lighter, cheaper, have higher screen resolutions and lower latencies, making them more comfortable to use and greatly improving the user experience. As a result, HMD are now at a point where they will slowly start to affect the way we consume digital content in our everyday lives.

One of the main issues of wearing an HMD is that they are very invasive, and hide completely the wearer's face, see Figure 1. In many cases, this is not an issue since the user is isolated in a purely individualistic experience. However, HMDs can also be used in social scenarios. One example can be collaborative 3D immersive games where two individuals play together and can still talk and see each other's faces. Another example is video-conferencing, where switching from traditional screens and cameras to 3D acquisition + HMDs can bring the possibility of viewing the other person and his surroundings almost as if he was really there physically. In these and other cases, not seeing the other person's face clearly damages the quality of the experience.

In this paper we propose to remove the HMD and replace it by the face of the wearer, using 3D graphics and computer vision techniques. Our method involves three separate steps: 1) building a 3D face model of the user from several photographs, 2) re-project the model into the video mimicking the person's head pose and facial expressions (estimated via HMD's tracking and face landmarking techniques respectively) and 3) smartly combine both images in a seamless way, keeping the visible parts of the face (e.g. mouth, chin) while replacing those hidden by the HMD (e.g. eyes, nose, etc.). Our method works at real-time speeds to ensure its applicability to video streaming scenarios.

One of the main features of our method is that it is able to recover the main facial expressions a user can portray. Using face landmark

estimation techniques, we train a system to learn the facial expressions of the user from the lower part of the face only, and use it to change the model accordingly at test time. The result is a face recovery where the upper part of the face changes in synchrony with the movement of the lower (visible) part, greatly improving the quality of the overall result.

## 2 Related work

To the best of our knowledge, we are the first to propose face reconstruction of a person wearing a HMD via on-line tracking, facial expression detection and user-specific 3D head models. However, our work shares common ground with general occluded face reconstruction, as well as with prior art in face transfer, which consists in swapping expressions across two different people.

**Face reconstruction:** [Hwang and Lee 2003] were among the first to tackle the problem of recovering partially occluded faces. Their appearance learning method was based on a 2D morphable model and faces were prototyped in a PCA-based projection of both shapes and textures, much like in the original AAM formulation [Cootes et al. 2001]. Similarly, [Mo et al. 2004; Yu and T. 2008; Hosoi et al. 2012] proposed to recover missing regions by exploiting the correlations between nearby regions in aligned shape space using other linear dimensionality reduction techniques such as FW-PCA or linear mixtures. Finally, [D.Lin and Tang 2007] proposed a Bayesian framework that also allows the automatic detection of face occlusions. These methods often use general image in-painting techniques [Bertalmio et al. 2000; Pérez et al. 2003] as a post-processing step to improve the final result.

These approaches assume that head pose and expressions are constant or known, that all faces were either previously aligned, that ground-truth facial landmarks are given and that occlusion mask is known both during training and testing. None of these assumptions hold in our scenario, hence the need for a novel method.

**Face transfer:** Another related application is that of swapping two faces, transferring a source face to a target face while conserving the original facial expression of the target. Some examples are [Vlasic et al. 2005; Bitouk et al. 2008; Cheng et al. 2009; Dale et al. 2011; Huang and De La Torre 2012]. These methods often use 3D morphable models [Blanz and Vetter 1999] of the face, 3D reconstruction [Chen and Medioni 1992; Baillard and Zisserman 2000; Faugeras and Luong 2004] and in-painting techniques similarly to what proposed in this paper. However, they deal with fully unoccluded faces, and many of the methods proposed cannot work when more than half of the face is occluded as is the case in our scenario. Moreover, they seldom run in real-time.

**HMD Face expression detection:** Also somewhat related is recent work by [Li et al. 2015], where a new HMD is developed that enables 3D facial performance-driven animation by adding strain sensors mounted on the foam liner of the headset and a head-mounted RGB-D camera to enhance the tracking in the mouth region. While we welcome any improvement to current HMD's, we argue that much can be done using a simple video camera and modern computer vision techniques, as shown in this paper.

One of the backbones of our method is face landmark estimation, which consists in locating the position of a sparse set of pre-defined 2D key-point landmark locations encoding shape (commonly including, for example, the corners of the eyes, mouth, and nose). How to robustly estimate these landmarks is a widely studied field in computer vision. We use the method described in [**?**] due to its robustness, speed and availability of code online, but any other could equally be used. Reviewing all existing face landmarking methods falls behind the scope of this paper due to tight space constraints.

## 3 Method

Figure 1 shows an outline of the proposed method. The first part, explained in Section 3.1, consists in building (off-line) a user specific model that includes a discrete representation of the user facial expressions. Then, the method performs real-time face reconstruction by tracking the HMD's position, detecting the user facial expression and in-painting the occluded pixels with the help of the projected model (see Section 3.2).

### 3.1 Building the user-specific model

**3D mesh building:** The first step is to build a textured 3D model of the user's head from a series of photographs of the user taken from varied view-points (5-10 normally suffice). This can be done using available free software such as [Autodesk Inc. ].

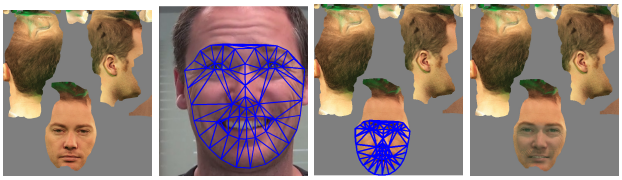
**Adding facial expressions:** The second step is to learn the array of expressions the user usually portrays when he acts naturally. We collect several images/videos of the user talking (without wearing an HMD) and we apply to each an automatic face landmarking method [**?**] to estimate the shape of the face. This results in $N$ training shapes $\mathcal{S} =< \mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_N >$, where $N$ is the total number of images collected. Each $\mathbf{S}$ encodes the $x$ and $y$ image positions of $L$ landmarks $\mathbf{S} =< \mathbf{X}, \mathbf{Y} >\in \mathbb{R}^{2L}$. A typical representation is Multi-Pie $L = 68$ landmark format [Gross et al. 2008].

The shapes are normalized to remove variations due to different face position and resolutions. Then, we cluster them using standard K-means algorithm with $K$ being the number of expressions we want to discover. This results in a partition of the data into K clusters, each containing $N_k$ images. Each cluster $\mathbf{C}_k$ is represented by the average of all the normalized shapes it contains:

$$\mathbf{C}_k \equiv< \bar{\mathbf{X}}, \bar{\mathbf{Y}} >\equiv \frac{1}{N_k} < \sum_{n=1}^{N_k} \mathbf{X}_n, \sum_{n=1}^{N_k} \mathbf{Y}_n > \quad (1)$$

The cluster centers $\mathcal{C} =< \mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_K >$ form our expression classifier. Once the clusters are created, for each we learn a texture representation which best summarizes its $N_k$ images, see Figure 2. First, we decompose the image following a standard delaunay triangulation on the landmarks, decomposing the image into a set of landmark-indexed triangles:

$$DT(\mathbf{S}) =< T_1, T_2, ..., T_T >, \quad (2)$$
where each $Tt = [< \mathbf{a}, \mathbf{b}, \mathbf{c} >\in [1..L] \wedge a \neq b \neq c]$



Model UV map | Cluster landmark triangulation | Model landmark triangulation | Transfer result

**Figure 2:** *Building expression-specific textures*

When projected into the image, the convex hull of each triangle $T_t$ contains a set of pixels. Rasterizing these pixels across all $N_k$ images result in a set of vectors:

$$\mathcal{P} =< \mathbf{P}_1, ..., \mathbf{P}_{N_k} > \in \mathbb{R}^M \quad (3)$$

$$\mathbf{P}_n =< raster(Image_n, T_1), ..., raster(Image_n, T_T) >$$

Each $\mathbf{P}$ has a fixed size $M$ (each triangle is warped across all $N_k$ images to have the same size). From these vectors, we apply dimensionality reduction techniques (e.g. PCA [Pearson 1901]) and keep only the first $Q$ principal components (whith $Q << N_k$), resulting in a set of $Q$ vectors $\mathcal{P}' =< \mathbf{P}'_1, \mathbf{P}'_2, ..., \mathbf{P}'_Q >$. Finally, our cluster-specific texture representative $\bar{\mathbf{P}}$ is the average pixel intentity of each pixel for all $Q$ vectors $\bar{\mathbf{P}} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{P}'_q$.

This process is repeated for each cluster, resulting in $K$ cluster-specific textures $\bar{\mathbf{P}}_1 ... \bar{\mathbf{P}}_K$. Finally, we also establish the position of the landmarks in the original UV Map of the textured model (manually). With this equivalence, we can transfer to the UV map our texture representative $\bar{\mathbf{P}}_k$ by warping and copying the triangle contents. Repeating this for each $K$ cluster results in $K$ expression-specific UV maps among which to choose.

### 3.2 Real-time face recovery

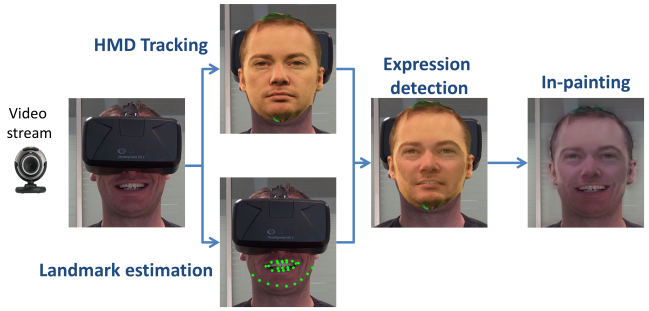



**Figure 3:** *Test time method outline*

Fig. 3 shows the steps involved at test time, we explain them below.

**HMD Tracking:** For a successful face recovery, we need to estimate the user's head pose in the camera frame. Most HMD provide an Inertial Measurement Unit (IMU) to assess the user's head rotations. The IMU suffices to adapt the video content to the user head movements but it remains a very raw and laggy estimation of the real head rotation. Moreover, it doesn't provide any information regarding translations. Therefore, video-based tracking techniques are needed. Two different methods have been developed, one for each HMD model tested.

**Oculus DK1:** For the DK1 we developed a fiducial marker-based solution. Four ArUco patterns [Garrido-Jurado et al. 2014] were fixed on the front and lateral parts of the Oculus rift, to allow a fixed video camera placed in front of the user to always catch at least one marker during important head rotations. During a typical tracking session, the 4 corners of each marker are detected making use of the image processing algorithms available in the ArUco library. The pose estimation is then obtained thanks to a Perspective-n-Point camera pose estimation algorithm such as ePnP [Lepetit et al. 2009] (assuming camera intrinsic parameters are known following a classical chessboard calibration).

**Oculus DK2:** The second version of the Oculus HMD comes with an embedded tracking system in addition to the previous IMU-based solution. It uses a layer of blinking infrared light-emitting diodes inlayed just behind the external surface of the helmet and a fixed infrared sensitive video camera installed in front of the user. This new system allows to obtain an accurate estimation of both the rotation and translation of the HMD.

However, the infrared camera cannot be set up to acquire color images and therefore cannot be used for the face recovery procedure. To overcome this limitation, a fixed regular color camera was used in combination with the infrared sensor. To obtain the user head pose in the color camera frame while exploiting the embedded infrared tracking system, a co-calibration procedure between the two sensors was carried out (not explained here due to space constraints).

**Landmark estimation:** In parallel to the HMD tracking, we apply the image-based face landmarking algorithm. We train a mouth-only model to avoid errors due to heavy occlusion in the other areas such as eyes, nose, etc. and use it to estimate the current shape $\mathbf{S}$ of the mouth in the video stream. We also enforce frame-to-frame temporal coherence via shape tracking techniques [**?**].

**Expression detection:** We classify the current frame as belonging to the expression-cluster with most similar mouth shape. We compute the distance between the current mouth/chin normalized shape estimate $\mathbf{S}$ and all the cluster centers $\mathcal{C} =< \mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_K >$. That which has minimal distance is considered to be the best match:

$$\arg\min_k \sum_{k=1}^{K} \|C_k - \mathbf{S}\|^2 \quad (4)$$

The 3D mesh is filled-in using the UV map created during training from the representative $\bar{\mathcal{P}}_k$ of the expression $k$ detected.

**In-painting:** Once the 3D expression-specific model is correctly aligned on top of the current video frame, in-painting operations are performed to achieve a seamless face recovery. We use the HMD position information to select which parts of the 3D model are kept and which removed (occlusion mask). To reduce differences in color between the texture in the 3D model and the current video frame, we first perform a statistical color transfer to alter the 3D model texture color so that it matches that of the target image (where $\mu$=average and $\sigma$ = standard deviation):

$$src = \left( \frac{\sigma_{target}}{\sigma_{src}} * (target - \mu_{src}) \right) + \mu_{src} \quad (5)$$

Then, we perform laplacian blending on the occluded region, merging pixels from the model, the video frame and the background (learnt off-line). The blending removes discontinuities at the borders and further polishes color differences, resulting in a seamless composition.

## 4 Results

Figure 4 shows some qualitative example results on three users, using two different HMD models (Oculus DK1 and DK2) each with a different tracking system (pattern-based and infrared-led). See also videos in Supp. Material. Please note that we cannot measure quantitative results due to lack of ground-truth (face is hidden naturally and not artificially). In all cases we used [$K = 6$, $Q = 10$] empirically found to give best results.

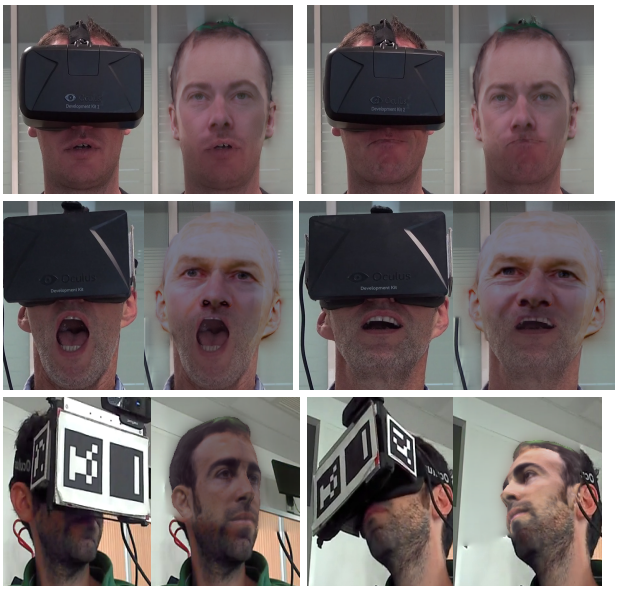


**Figure 4:** *Some example results. See also videos in Supp. material.*

These results show the potential of our method. Compared to the original, occluded input, the person is clearly recognizable, and while the replacement is far from perfect, it is convincing. When the expression is correctly detected, the emotion of the user is clearly conveyed. Our method, implemented in C++ with parts in GPU (GLSL) runs at 30 frames per second in a standard desktop PC.

We consider these results very encouraging considering the difficulty of the task (working with faces occluded by more than 80% is no easy task). For future work, we will focus on improving the model building and the expression detector. The former can probably benefit from the use of 3D morphable models and blendshapes. The latter remains a challenging task, since some expressions resemble each other very closely when judged solely from the mouth (*e.g.* doubt and fear). However, we think that results can be improved by adding appearance features and leveraging large human-expression public datasets as general training examples to complement the user-specific ones.

## References


AUTODESK INC. 123d catch app. http://www.123dapp.com/catch.

BAILLARD, C., AND ZISSERMAN, A. 2000. A plane-sweep strategy for the 3d reconstruction of buildings from multiple images. *International Archives of Photogrammetry and Remote Sensing 33*, B2; PART 2, 56–62.

BERTALMIO, M., SAPIRO, G., CASELLES, V., AND BALLESTER, C. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., 417–424.

BITOUK, D., KUMAR, N., DHILLON, S., BELHUMEUR, P., AND NAYAR, S. K. 2008. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph. 27*, 3, 39:1–39:8.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*.

BURGOS-ARTIZZU, X., PERONA, P., AND DOLLAR, P. 2013. Robust face landmark estimation under occlusion. In *International Conference in Computer Vision (ICCV)*, IEEE.

CHEN, Y., AND MEDIONI, G. 1992. Object modelling by registration of multiple range images. *Image and vision computing 10*, 3, 145–155.

CHENG, Y.-T., TZENG, V., LIANG, Y., WANG, C.-C., CHEN, B.-Y., CHUANG, Y.-Y., AND OUHYOUNG, M. 2009. 3d-model-based face replacement in video. In *SIGGRAPH '09: Posters*.

COOTES, T., EDWARDS, G., AND TAYLOR, C. 2001. Active appearance models. *PAMI 23*, 6, 681–685.

DALE, K., SUNKAVALLI, K., JOHNSON, M. K., VLASIC, D., MATUSIK, W., AND PFISTER, H. 2011. Video face replacement. *ACM Trans. Graph. 30*, 6, 130:1–130:10.

D.LIN, AND TANG, X. 2007. Quality-driven face occlusion detection and recovery. In *CVPR*.

FAUGERAS, O., AND LUONG, Q. 2004. *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications*. MIT Press.

GARRIDO-JURADO, S., NOZ SALINAS, R. M., MADRID-CUEVAS, F., AND MARÍN-JIMÉNEZ, M. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition 47*, 6, 2280 – 2292.

GROSS, R., MATTHEWS, I., COHN, J., KANADE, T., AND BAKER, S. 2008. Multi-pie. In *FG*.

HOSOI, T., NAGASHIMA, S., KOBAYASHI, K., ITO, K., AND AOKI, T. 2012. Restoring occluded regions using fw-pca for face recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*.

HUANG, D., AND DE LA TORRE, F. 2012. Facial action transfer with personalized bilinear regression. In *Computer Vision–ECCV 2012*. Springer, 144–158.

HWANG, B.-W., AND LEE, S.-W. 2003. Reconstruction of partially damaged face images based on a morphable face model. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 3, 365–372.

LEPETIT, V., F.MORENO-NOGUER, AND P.FUA. 2009. Epnp: An accurate o(n) solution to the pnp problem. *International Journal Computer Vision 81*, 2.

LI, H., TRUTOIU, L., OLSZEWSKI, K., WEI, L., TRUTNA, T., HSIEH, P.-L., NICHOLLS, A., AND MA, C. 2015. Facial performance sensing head-mounted display. In *SIGGRAPH*.

MO, Z., LEWIS, J., AND NEUMANN, U. 2004. Face inpainting with local linear representations. In *BMVC*.

OCULUS VR. Oculus rift dk2. https://www.oculus.com/dk2/.

PEARSON, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*.

PÉREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. In *ACM Transactions on Graphics (TOG)*, vol. 22(3), ACM, 313–318.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, vol. 24(3), ACM, 426–433.

YU, D., AND T., S. 2008. Using targeted statistics for face regeneration. In *IEEE International Conference on Automatic Face and Gesture Recognition*.