Assessing Facial Expressions in Virtual Reality Environments

Catarina Runa Miranda and Verónica Costa Orvalho
Instituto de Telecomunicações, Universidade do Porto, Porto, Portugal

Keywords: Facial Motion Capture, Emotion and Expressions recognition, Virtual Reality.

Abstract: Humans rely on facial expressions to transmit information, like mood and intentions, usually not provided by the verbal communication channels. The recent advances in Virtual Reality (VR) at consumer-level (Oculus VR 2014) created a shift in the way we interact with each other and digital media. Today, we can enter a virtual environment and communicate through a 3D character. Hence, to the reproduction of the users' facial expressions in VR scenarios, we need the on-the-fly animation of the embodied 3D characters. However, current facial animation approaches with Motion Capture (MoCap) are disabled due to persistent partial occlusions produced by the VR headsets. The unique solution available for this occlusion problem is not suitable for consumer-level applications, depending on complex hardware and calibrations. In this work, we propose consumer-level methods for facial MoCap under VR environments. We start by deploying an occlusions-support method for generic facial MoCap systems. Then, we extract facial features to create Random Forests algorithms that accurately estimate emotions and movements in occluded facial regions. Through our novel methods, MoCap approaches are able to track non-occluded facial movements and estimate movements in occluded regions, without additional hardware or tedious calibrations. We deliver and validate solutions to facilitate face-to-face communication through facial expressions in VR environments.

1 INTRODUCTION

In the last two decades, we lived a revolution of global digital interactions and communication between humans (Jack and Jack, 2013). We erased geographic barriers and started communicating with each other through phones, computers and, more recently, inside virtual environments using Virtual Reality (VR) headsets. Oculus VR company was the responsible by bringing this hardware to consumer-level making this way of interaction more appealing to common users (Oculus VR 2014). However, VR communications remain a challenge. Human communication strongly rely on a synergistic combination of verbal (e.g. speech) and non-verbal (e.g. facial expressions and gestures) signals between interlocutors (Jack and Jack, 2013). Past communication technologies, like phones and computers, adopted the image stream (e.g. webcams) coupled with speech to transmit both signals creating more realistic and complete experiences (Lang et al., 2012). In VR scenarios, we cannot use image stream since we are interacting with the virtual world embodied in 3D characters (Biocca, 1997; Slater, 2014). As result, the demand for on-the-fly algorithms for 3D characters animation and interaction is even higher. Ahead of unlocking both communication channels (i.e. verbal and non-verbal), the believable animation of 3D characters using user's movements enhance the three components of the sense of embodiment in VR environments: self-location, agency and body ownership (Biocca, 1997; Kilteni et al., 2012). Even with technological advances in Computer Vision (CV) and Computer Graphics (CG), the reproduction of human's facial expressions as facial animation of 3D characters is still hard to achieve (Pighin and Lewis, 2006). To automatise facial animation, facial Motion Capture (MoCap) has been widely used to trigger animation (Cao et al., 2014; von der Pahlen et al., 2014; Cao et al., 2013; Li et al., 2013; Weise et al., 2011). However, these approaches are not suitable for consumer-level VR applications, requiring or expensive setups (von der Pahlen et al., 2014), manual complex calibrations (Cao et al., 2013; Li et al., 2013; Weise et al., 2011) or do not support the persistent partial occlusion of the face produced by VR headsets (Cao et al., 2014).

To overcome the tracking problem created by persistent partial occlusions, Li *et al.* (Li et al., 2015) proposed a hardware based solution using a RGB-D camera for capture and strain gauges (i.e. flexible metal foil sensors) attached to VR headset to measure the upper face movements that are occluded. But

Miranda, C. and Orvalho, V.
Assessing Facial Expressions in Virtual Reality Environments.
DOI: 10.5220/0005716604860497
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3: VISAPP*, pages 488-499
ISBN: 978-989-758-175-5

again, this approach is not suitable for general user. It requires a complex calibration composed by hardware calibration to user and a blendshapes calibration to trigger animation. At the moment, this is the unique on-the-fly facial animation with MoCap solution compatible to VR environments.

Contributions: This work delivers and validates consumer-level real-time methods for: (i) facial MoCap method for persistent partial occlusions created by VR headsets and (ii) facial expressions prediction algorithms of occluded face region using movements tracked in non-occluded region. Compared to literature, we reduce user-dependent calibration and hardware requirements, requiring only a common RGB camera for capture. Our methods make current facial MoCap approaches compatible to VR environments and enable the extraction of key facial movements of bottom and upper face regions. The movements tracked and emotions detected can be combined to: trigger on-the-fly facial animation, enabling non-verbal communication in VR scenarios; as input for emotion-based applications, like emotional gaming (e.g. Left 4 Dead 2 by Valve).

2 BACKGROUND

In this section, we aim to study the literature regarding two different topics: (i) facial MoCap solutions for persistent partial occlusions created by VR Head Mounted Displays (HMD) and (ii) partial occlusions impact in facial expressiveness. The first topic presents state of the art facial MoCap solutions to overcome the persistent occlusions' issue. Then, in (ii), we explore how these occlusions restrict face-to face communication and their impact in face expressiveness. By the end, we search for a connection between occluded and non-occluded facial parts used as guide for methodology definition.

2.1 Persistent Partial Occlusions: A Today's Problem

In literature, we are able to find several promising solutions for real-time automatic facial MoCap (Cao et al., 2014; von der Pahlen et al., 2014; Cao et al., 2013; Li et al., 2013; Weise et al., 2011). However, the arise of VR commercial approaches of consumer-level HMD's (Oculus VR 2014), raised a new issue: the real-time automatic tracking of faces partially occluded by hardware (i.e. persistent partial occlusions of face) (Slater, 2014). Current MoCap approaches adopt model-based trackers, which produce cumulative errors in presence of persistent partial occlusions (Cao et al., 2014). Therefore, due to the absence of VR devices in mass-market, this issue was almost ignored for years. This resulted in a lack of technological solutions for face-to-face communication for VR environments. Only in 2015, Li *et al.* (Li et al., 2015) highlighted this problem and proposed a hardware based tracking solution. This solution uses an RGB-D camera combined with eight ultra-thin strain gauges placed on the foam liner for surface strain measurements to track upper face movements, occluded by the HMD. The first limitation of this approach is the long initial calibration required to fit the measures to each individual's faces using a training sequence of FACS (Ekman and Friesen, 1978). Also, in subsequent wearings by the same person, a smaller calibration is needed to re-adapt the hardware measures. This training step allows the detection of user's upper and bottom face expressions and activate a blendshape's rig containing the full range of FACS shapes (Ekman and Friesen, 1978). Besides the manipulation complexity, the solution also presents drifts and decrease of accuracy due to variations in pressure distribution from HMD placement and head orientation. As consequence, HMD straps positioning influence eyebrows' movement detection (Li et al., 2015). Li *et al.* solution is currently the only one available to overcome the persistent partial occlusions issue, making this an open research topic in CV algorithms for facial MoCap.

2.2 Partial Occlusions and Expressiveness

Everyday, humans' communication use facial expressions and emotions to transmit and enhance information not provided by speech (Lang et al., 2012). Even through technology, we always search for a way to use the non-verbal communication channel. As example, using video stream of our faces; virtual representations, like *emotion smiles*, cartoons or 3D characters with pre-defined facial expressions, etc. Understanding facial expressions and improve their representation in 3D characters is one of the key challenges of CG and plays an important role in digital economy (Jack and Jack, 2013). This role is even more relevant now, with recent advances in VR communications at consumer level (Biocca, 1997). *But how can we use the common solutions of facial animations, like MoCap, if user's face is occluded? Are we able to represent faces using information only from bottom of the face?* To answer these questions, we make a literature overview regarding several face regions impact in non-verbal communication. The goal is to understand how a partial occlusion of the face affects communication. We also researched for a relationship between occluded and non-occluded facial parts through emotion-based and biomechanics studies. This information was used to build one of this work hypothesis.

In a study about face perception (Fuentes et al., 2013), we concluded that humans have independent shape representations of upper and bottom parts of the face. Similar conclusions are found in emotion perception's literature, where mouth and eyes play different roles (Eisenbarth and Alpers, 2011; Lang et al., 2012; Bombari et al., 2013). In (Eisenbarth and Alpers, 2011; Bombari et al., 2013) it is shown that according to the emotion detected participants used information from eyes, or mouth or both. More precisely, in happy expressions participants used information from the mouth; for sad and angry, from eyes; and to fear and neutral, both mouth and eyes are used. For additional information about non-verbal communication, we forward the reader to (Lang et al., 2012). Taking these statements into account, if we occlude certain region of the face, face-to-face communication is affected and we may not be able to decode expressions properly. Subsequently, the tracking of only certain facial regions, like mouth, is not enough for emotion recognition, for proper communication and to generate believable facial animation of 3D characters.

From the biomechanical point of view, we know that facial muscles work synergistically to create expressions. The muscles interweave with one another, being difficult to decode their boundaries, since their terminal ends are interlaced with other muscles. A detailed research about facial anatomy and biomechanics can be accessed at Chapter 3 of the book *Computer Facial Animation* (Parke and Waters, 1996). Several studies in CG applied the biomechanical approach to create coding systems. These coding systems parameterize human face enabling a faster generation of facial expressions in 3D characters (Ekman and Friesen, 1978; Pandzic and Forchheimer, 2003; Magnenat-Thalmann et al., 1988). Although, they do not provide a clear solution for facial expressions estimation constrained to certain regions of the face. Furthermore, the definition and prediction of facial expressions is even harder when the diversity of facial expressions is considered. Scott McCloud (McCloud, 2006) explains the infinite possibilities of facial expressions combinations (i.e. the way mixing any two of universal emotions can generate a third expression, which, in many cases, is also distinct and recognizable enough to earn its own name) (McCloud, 2006).

Then, analyzing literature, we are able to attain that occlusions generated by VR devices affect communication and using only the information of non-occluded regions is not enough to animate a 3D character. However, biomechanics and facial animation coding systems show a connection between the different facial regions and how diverse and complex is the world of possible expressions. Using these statements, we describe a novel methodology to overcome occlusions problem of facial MoCap and then, to assess facial expressions using non-occluded face information.

3 METHODOLOGY

The literature overview of previous section allowed us to formulate the following hypothesis:

to create a method to estimate facial expressions of upper face and emotions using only bottom face's movements.

Therefore, we deliver VR consumer-level methods that:

- overcomes the persistent partial occlusions issue in MoCap, making possible the bottom face's movements tracking;
- recognizes universal emotions, plus neutral (Ekman and Friesen, 1975; Jack and Jack, 2013) using bottom face's movements;
- estimates upper face's movements (i.e. eyebrows movements) using information tracked from bottom part of the face.

Figure 1 shows the connection between our VR methods. We start by presenting a method to make generic MoCap systems compatible to persistent partial occlusions produced by VR headsets. Then, applying this algorithm, we are able to track properly the bottom face's features and use them to develop methods that predict the following facial expressions: (i) universal emotions, plus neutral (Ekman and Friesen, 1975; Jack and Jack, 2013) and (ii) eyebrows movements. Combining aforementioned methods, we make possible the MoCap of upper and bottom face movements and estimation of facial emotions under persistent partial occlusions created by VR headsets.

As setup, we suggest the usage of a Head Mounted Camera (HMC) combined with the VR HMD (see Figure 2). At first, we justify the adoption of HMC as capture hardware: When the user is inside the VR environment he is not aware of the space around him. The VR devices precisely substitute the user's sensory input and transform the meaning of their motor outputs with reference to an exactly knowable alternate reality (Slater, 2014). Hence, the user moves and

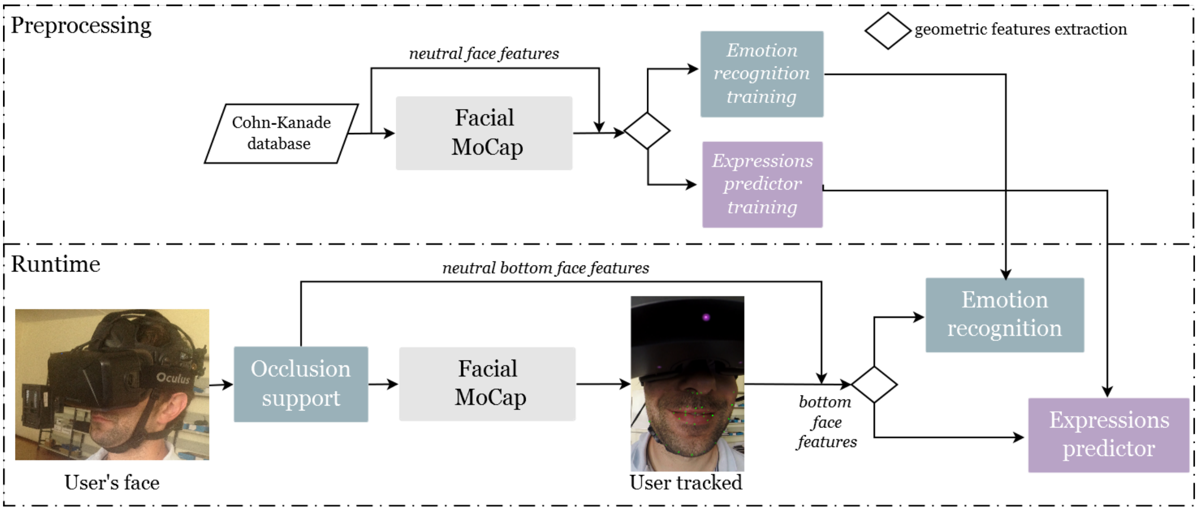

Figure 1: VR methods' framework.

reacts to impulses from VR environment. If we want to capture his face, we have to attach a capture device (i.e. camera) to his body and the device should follow user's movements (see HMC on Figure 2). It is not possible to use a static camera, because the user is not going to be able to place himself in a position proper for capture. A similar setup was also proposed by Li *et al.* (Li et al., 2015), but we removed the strain sensors.

In the next subsections, we provide a complete description of the VR methods.

3.1 VR Persistent Partial Occlusions: A Novel Method

To deploy our occlusion support method for facial MoCap, we used the following statement: we know the kind of occlusion created by HMD, so we know which part of the face is occluded. We also know that MoCap algorithms fail in these situations because they use a face model. When the face is occluded this model starts not to fit since there is not a full face being captured. As a solution, we use the knowledge that the region occluded is the upper part of the face to "re-create" the whole face.

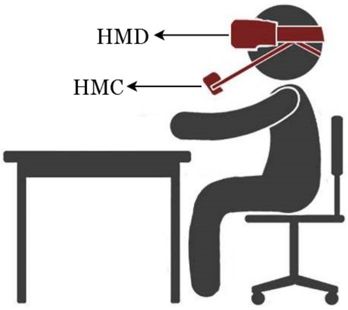

Figure 2: VR setup definition.

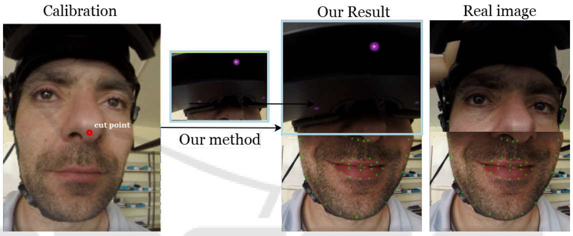

Figure 3: VR method: Persistent partial occlusions. From left to right: calibration image without VR HMD; our method uses cut point (red circle) to cut image an overlay at subsequent images: at left, what facial MoCap method see is a full face and, at right, the real image.

Our novel method overlays the upper part of the face captured on a neutral pose during calibration. Firstly, we assume that the higher visible point of the face is the nose and define it as cut point (i.e. this point can be changed to fit the occlusion created by certain HMD). Then, we detect the cut point with the MoCap and we cut the upper part of the calibration image (i.e. frame streamed) from the nose up, and use it to overlay to all the next camera/video frames. Hence, now the occluded part of the face is replaced with a static neutral face. The MoCap system is now able to detect the features in the combined half static/ half expressive face (see Figure 3). We ensure a proper re-creation of a face since we use a HMC that removes the user's head movements, i.e. user's face is in the same position during calibration and next streamed images.

3.2 VR Assessing Facial Expressions

During the development of VR facial expressions method, we applied face features and machine learning know-how from our past real-time emotion recognition research (Loconsole et al., 2014). In this novel method, we set the following goals: real-time emotion recognition of universal emotions (Ekman and Friesen, 1975) and upper face expressions prediction under VR scenarios. We aim to track facial expressions ahead of only emotions, in order to get a wide change of facial expressions and better cover and representation of the diversity of faces (McCloud, 1993). In opposition to the emotion classification method (Loconsole et al., 2014), where we needed to reduce the number of features tracked, in VR scenarios we have to maximize the information tracked in the bottom part of the face. Therefore, the feature extraction method should be able to retrieve enough information to allow an accurate prediction of facial expressions by the machine learning algorithm.

As a solution, we propose to use all the features tracked of bottom face region (see Figure 4 blue rectangle) and apply a geometrical features extraction algorithm. This algorithm is defined as the Euclidean distance between neutral face features (stored during calibration step of previous persistent partial occlusions method) and current frame (i.e. instant in time) features. Summarizing, to each feature tracked p in certain instant i, we calculate the distance $D(p_i, p_c)$:

$$D(p_i, p_c) = \sqrt{\frac{((p_i(x) - p_c(x))^2 + (p_i(y) - p_c(y))^2}{\|p_i - p_c\|}}$$

,where:

p_i is the 2D bottom face feature p at the instant i in time;

p_c is the 2D bottom face feature p of neutral expression captured during calibration;

$\|p_i - p_c\|$ is the norm between p_i and p_c in Cartesian space.

Since the occlusion produced varies according to VR headset used, we also created machine learning models to assess facial expressions using the bottom face features information including and excluding nose features. The bottom face features without nose feature can be used by the different kinds of HMD, since the nose region is the one affected by the device size.

To create the machine learning models to predict the emotions and upper face expressions, we used the Cohn-Kanade (CK+) database (Lucey et al., 2010). CK+ database contains posed and spontaneous sequences from 210 participants (i.e. cross-cultural adults of both genres). Each sequence starts with a neutral expression and proceeds to a peak expression. This sequences are FACS coded and emotion labeled. The transition between neutral and a peak expression allowed us to detect spontaneous expressions and not only pure full expressions.

To implement the algorithms, we adopted a GPU version of Random Forest (Breiman, 2001) of OpenCV (OpenCV, 2014) to generate respective machine learning models for real-time prediction. As facial MoCap testing approach, we deployed the Saragih *et al.* (Saragih et al., 2011) system. (see Figure 4 tracking landmarks in green).

3.2.1 VR Emotion Recognition: Novel Method

As preprocessing stage, we create the Random Forests model that is used to predict emotions in real-time (Loconsole et al., 2014). To build the model for emotion classification, to each database's sequence we applied the facial MoCap method and extracted bottom face features. Using the first frame of the sequence as neutral expression, to subsequent frames in the sequence, we calculate the distance $D(p_i, p_c)$, between bottom face features of current frame and neutral expression's frame. Thus, to train the machine learning model for emotion recognition we used aforementioned geometrical extraction algorithm: distance $D(p_i, p_c)$ of bottom face's features of each frame. As response value, to each distance calculated, we used respective CK+ emotion label (see Figure 4 blue processes).

As observed in the Figure 2, in runtime, we apply once our occlusions support method and store neutral face features. This step is only execute one time per user. After, in runtime, the adapted facial MoCap system delivers bottom face's movements and distance $D(p_i, p_c)$ is calculated to each feature p. The group of distances are used as input in the Random Forests classifier that predicts the user's emotion represented by that distances and respective accuracy's percentage.

3.2.2 VR Facial Expressions Predictor: Novel Method

To build the upper face expressions model, we also applied the distance of neutral and expression bottom face features as geometric extraction algorithm. However, we have to define the movements that we wanted to predict in order to create specific tags to the training process. For simplicity, we set as upper face expressions the prediction of eyebrows movements, i.e. the detection if eyebrows are going up or down, and the ”how much” they are moving compared to a neutral position. This last parameter is measured as a percentage of movement up/down compared to neutral expression. Similarly to assumption made in

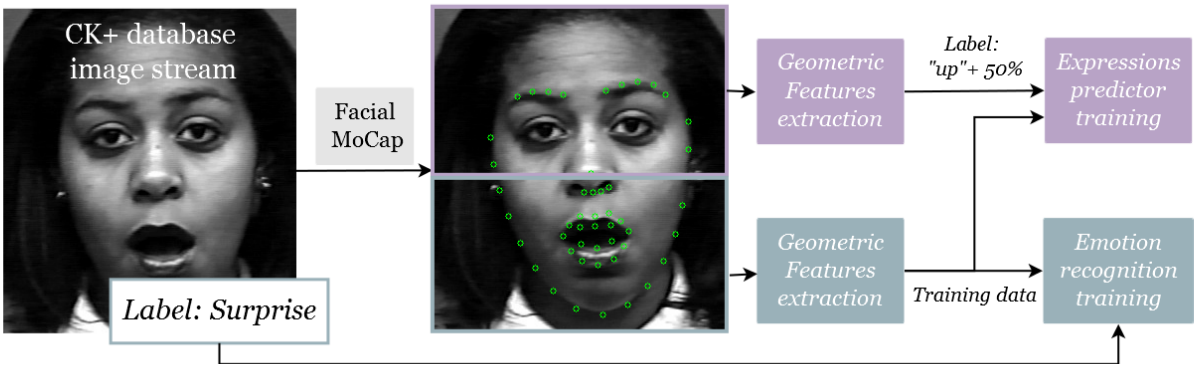
CK+ database image stream
Label: Surprise
Facial MoCap
Geometric Features extraction
Label: "up"+ 50%
Expressions predictor training
Geometric Features extraction
Training data
Emotion recognition training

Figure 4: VR methods: Expressions predictor training (purple) and Emotion predictor training (blue) with CK+ database.

(Fuentes et al., 2013), we assume symmetry of the eyebrows movements. To define the tags, we calculated the Euclidean distance $D(p_i, p_c)$ between neutral position of eyebrows and the expression positions in the other frames of the sequence. If the average of the eyebrows features indicated that they are going up, we tagged ”up”; the opposite if the eyebrows went down we tag ”down” (i.e. we used image coordinate system, so this distance was negative when eyebrows go up and vice-versa). Simultaneously to each frame of the sequence tagged we saved the percentage of movement compared to neutral position (up or down). As result, to each frame of the sequence of each participant in CK+ database we tagged: eyebrows ”up” or ”down”, plus percentage of movement. In Figure 4 with purple processes, the reader can observe an example of method's framework.

At preprocessing stage, we trained two Random Forests models with the same input data: the distances $D(p_i, p_c)$ between neutral and current bottom face features; but using one of the following response values:

- ”up” and percentage of movement, if eyebrows are rising
- ”down” and percentage of movement, if eyebrows are descending

, to each frame of each sequence of CK+ database.

Since we are using a GPU approach of the classifier, with high computational performance, to maximize the prediction accuracy of eyebrows movements, we trained two models: one to predict the rise movement and, other, to predict the opposite. In runtime, we apply the defined geometrical features extraction to the bottom face's features tracked by the adapted MoCap. The extracted features are used as input in both Random Forests classifiers, to retrieve one of the predictions:

1. **eyebrows ”rising”** and percentage of movement;
2. **eyebrows ”descending”** and percentage of movement.

Since we are using two different classifiers, there is a probability of confusion of both models return simultaneously an ”up” and ”down” movement. As a solution, our method compares the accuracies of prediction from the two classifiers' predictions, and the result delivered is the one with higher accuracy.

4 RESULTS AND VALIDATION

In this section, we show the results and statistical validation of the methods proposed. Statistical analysis was performed using R software (R Core Team, 2013).

4.1 VR Persistent Partial Occlusions

To test our occlusions method, we applied it to Saragih *et al.* (Saragih et al., 2011) and Cao *et al.* (Cao et al., 2014) MoCap systems (see Figures 5 and 6, respectively). At the Figure 7, we test a generic partial occlusion created by a piece of paper.

As observed in the Figures 5, 6 and 7, our occlusion-support method adapts to MoCap systems making them compatible to persistent partial occlusions. The ”paper” test case represented a generic occlusion created by a random VR device. As conclusion, our method is not only adaptable to MoCap, but it could be also used to generic partial occlusions created by different VR HMD's.

4.2 VR Assessing Facial Expressions

We divided the validation of our prediction methods in two steps: (i) statistical validation and (ii) visual validation.

To validate statistically our machine learning classifiers we adopted a k-Fold Cross Validation (k-Fold

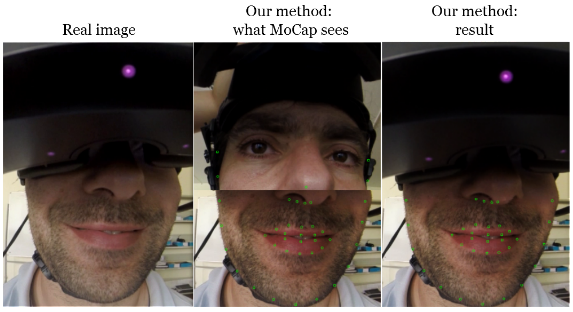

Figure 5: VR method results: Persistent Partial Occlusions method applied to Saragih *et al.* (Saragih et al., 2011) MoCap. The real image (left), our method result and what MoCap processes (middle) and final result from our method (right).

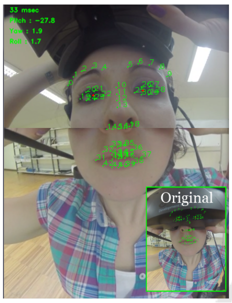

Figure 6: VR method results: Persistent Partial Occlusions method applied to Cao *et al.* (Cao et al., 2014) MoCap.

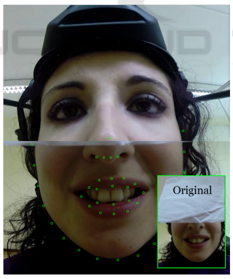

Figure 7: VR method results: Persistent Partial Occlusions method applied to Cao *et al.* MoCap algorithm (Cao et al., 2014) to overcome a general occlusion created by a piece of paper.

CRM) with k=10 (Rodriguez et al., 2010). The k-Fold CRM, after iterating the process of dividing the input data in *k* slices for *k* times, trains a classifier with *k-1* slices. The remaining slices are used as test sets on their respective *k-1* trained classifier, allowing us to calculate the accuracy of each one of the *k-1* classifiers. The final accuracy value is given by the average of the *k* calculated accuracies. Though, to each method we analyze k-Fold CRM accuracy to the methods under different scenarios. We highlight that this validation procedure ensures that the test dataset is not the same of the training dataset. Therefore, prediction accuracies are not calculated with test data contained in the training dataset.

Furthermore, we provide a statistical analysis of sensitivity versus specificity and positive versus negative predictive value (i.e. pred. in Tables) (Parikh et al., 2008). The sensitivity measures the performance of the classifier in correctly predicting the actual class of an item, while specificity measures the same performance but in not predicting the class of an item that is of a different class. Summarizing, sensitivity and specificity measure the true positive and true negative performance, respectively. We added the positive and negative predictive value analysis because these values reflect the probability that a true positive/true negative is correct given knowledge about the prevalence of each class in the data analyzed.

By the end of this section, we validated visually our VR methods regarding: occlusions, emotion and facial expressions prediction. The visual validation data was acquired in our laboratory and is not part of the training dataset (learning made CK+ database). The visual data was not acquired with HMC, but we asked to the participants to avoid extreme head movements. As result, we were able to test our VR method of occlusion-support and the facial expressions methods simultaneously.

4.2.1 VR Emotion Recognition

Using the k-Fold CRM, we executed a method's validation to two emotion recognition scenarios: (i) six universal emotions of Ekman and Friesen (Ekman and Friesen, 1975), plus neutral; (ii) four universal emotions of Jack (Jack and Jack, 2013), plus neutral. The six universal emotions (Ekman and Friesen, 1975) are the commonly used and accepted by literature studies. However, recent advances in psychology of the emotions show that these emotions are not reproducible throughout different cultures. The non-universality of Ekman's emotions is explored by the survey (Jack and Jack, 2013). This complete study defends that only a subset of the six ”universal” emotions is universally recognized, i.e. Joy/Happy, Surprise, Anger and Sad/Sadness. This subset excludes fear and disgust, since these emotions present low recognition cross-culturally being biologically adaptive movements from the emotions surprise and anger, respectively (Jack and Jack, 2013).

Therefore, the Table 1 shows the k-Fold CRM accuracies to the two scenarios.

In the Table 1, we observe an increase of the accuracy detection when recognizing four emotions, compared to six emotions classification. This result is not surprising, since we are reducing the number of emo-

Table 1: k-Fold CRM Accuracy comparison to scenario (i) and to the scenario (ii). Results in percentage (%).

Emotions	k-Fold Accuracy (%)	95% Confidence Interval
Six (Ekman and Friesen, 1978)	64.80	[61.72;67.79]
Four (Jack and Jack, 2013)	69.07	[65.59;72.40]

tions predicted. In addition, we detect that the bottom features of the face allow a weak recognition of face emotions, resulting in accuracies lower than 70%.

More in detail, we report in the Tables 2 and 3, a statistical analysis of each emotion recognition obtained with Random Forests classifier to scenario (i) and (ii), respectively.

Both statistical analysis resulted in a p-value lower than $2.2 \times e^{-}16$ to a significance level of 5%, which validates our method's hypothesis: classifying the six/four universal emotions using bottom of face features tracking. Specifically, to scenario (i) at the Table 2, we observe an overall low sensitivity to emotions classified (with exceptions to Joy/Happy and Neutral). The opposite is observed to specificity. This indicates that the method does not have high accuracy to detect a certain class, however, does not predict incorrectly. The predictive values weighted using information about the class prevalence in population, show an overall increase of accuracy for true positive and maintain to negative. Therefore, as example to Surprise, despite our classifier only being able to positively identify surprise in 59.40% of the time there is a 71.82% chance that, when it does, such classification is correct. Looking to Table 3, compared to previous results of scenario (i) at Table 2, we observe an increase of sensitivity, while maintaining an high accuracy of specificity. In general, the same is observed in positive and negative predictive values. This is expected, since decreasing the number of classes of emotions will decrease the degree of confusion that lead to a better split between classes, resulting in a better emotion recognition method. These results confirm the statement of Background section, i.e. bottom face features provide incomplete information about face expression of emotions. Though, our method presents better performance when four universal emotions (Jack and Jack, 2013) are classified.

4.2.2 VR Facial Expressions Predictor

To analyze and validate the VR facial expressions predictor, we executed the k-Fold cross-validation to the classifier eyebrows "rising" and to classifier eyebrows "descending". Taking into account the variance of nose tracking with the type of HMD used, we propose to study the influence of tracking these features (subset *S1*) and not tracking the nose features (subset *S2*) in the prediction of eyebrows' movements. Average K-Fold CRM accuracies and respective confidence intervals can be accessed in the Table 4.

In the Table 4, we observe a small decrease of accuracy when the nose features tracking is removed. Although, the confidence intervals show that this decrease is only significant in eyebrows "up" detection. Our method allows an high performance of eyebrows "up" estimation (at least, 85%) compared to eyebrows "down" estimation (at least, 66%). The different results arise from the fact that we are using an emotion database for training, where there is more data describing the "rising" movement than the opposite (i.e. only anger and sadness emotions usually present this facial expression behavior (Ekman and Friesen, 1978)).

Similarly to emotion recognition method, we present the statistical analysis of sensitivity/specificity and positive/negative predictive values to both eyebrows movements using the subsets *S1* and *S2*.

Both p-values of further analysis are lower than the significance level (i.e. p-value equal to $2.2 \times e^{-}16 < 0.05$). Therefore, both methods are suitable for eyebrows movement estimation using bottom face's movements. Table 4 shows that the method is able to classify the eyebrows "up" movement accurately, with exception for specificity using the subset *S2*. So, the removal of nose features tracking leads, essentially, to a decrease in accuracy of the classifier in not giving incorrect predictions. However, when we take in to account the prevalence of the class in population, the overall accuracy of prediction to both positive and negative values increase, presenting values above 84.04%.

Table 6 contains the statistical analysis to the prediction of eyebrows "descending" movement with (*S1*) and without *(S2)* nose features tracking.

Observing the Table 6, we observe that our method predicts correctly the "descending" movements of the eyebrows, at least, 73.18% of the time and does not predict incorrectly this movements in at least, 63.97% of the time. The lower values are obtained to the subset *S2*, however, the differences between subsets performance are not significant. Similar behavior is beheld taking into account the prevalence of the class in the population. The positive/negative predictive values are not significantly different between sensitivity/specificity. As expected by previous k-Fold CRM results, prediction of the

Table 2: Statistical Analysis of scenario (i) - Results in percentage (%).

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Sensitivity	53.15	39.44	26.09	81.29	12.70	59.40	90.80
Specificity	86.55	97.70	95.84	95.17	99.13	96.35	85.39
Positive pred.	40.21	57.14	39.34	75.90	50.00	71.82	75.51
Negative pred.	91.56	95.40	92.62	96.45	94.31	93.81	94.92

Table 3: Statistical Analysis of scenario (ii) - Results in percentage (%).

	Anger	Joy	Sadness	Surprise	Neutral
Sensitivity	75.50	77.85	13.80	68.75	80.09
Specificity	76.16	95.14	99.07	98.39	91.34
Positive pred.	45.06	81.46	66.67	88.51	80.44
Negative pred.	92.31	94.00	89.52	94.59	91.16

Table 4: k-Fold CRM Accuracy comparison facial expressions assessed (Eyebrows Up or Down) with subset *S1* and *S2*. Results in percentage (%).

Eyebrows movements	k-Fold Accuracy(%)	95% Confidence Interval
Up *S1*	91.47	[89.76;92.98]
Up *S2*	87.02	[84.97;88.89]
Down *S1*	70.63	[67.99;73.18]
Down *S2*	69.13	[66.40;71.76]

Table 5: Eyebrow Up prediction - Statistical Analysis to subsets *S1*. Results in percentage (%).

Eyebrows Up	*S1*	*S2*
Sensitivity	97.34	96.27
Specificity	71.79	59.18
Positive pred.	92.04	87.65
Negative pred.	92.31	84.06

Table 6: Eyebrow Down prediction - Statistical Analysis to subsets *S1*. Results in percentage (%).

Eyebrows Down	*S1*	*S2*
Sensitivity	77.13	73.18
Specificity	62.73	63.97
Positive pred.	71.57	72.09
Negative pred.	69.28	65.23

”descending” movement presents lower performance compared to prediction of the opposite movement. Again, this result occurred due to the low prevalence of the ”down” class in population. This statement is confirmed by the lower influence shown in positive and negative predictive values when compared to sensitivity and specificity, respectively.

Summarizing, our methods of facial expressions prediction are suitable for the estimation of eyebrows movements using features from the bottom of the face, specially in estimation of the ”rising” movement. This conclusion corroborates the hypothesis of this work: our results traduce a connection between bottom and upper face behaviors.

4.2.3 VR Assessing Facial Expressions: Visual Validation

Applying the methods to videos where the participants expressed emotions (Ekman and Friesen, 1975), we are able to check visually the performance of the methods: occlusions support, emotion recognition and expressions prediction. We chose a non-VR scenario in order to verify if the upper face movements and emotions predicted (using only bottom face's movements) match the original facial expressions. Results can be observed in the Figures 8, 9, 10 and 11.

Looking throughout the Figures, we verify that our occlusion method is able to ”re-create” the face even not using a HMC. Regarding emotion recognition using only the facial features (green dots), in the Figure 8, 9 and 10, we show three examples of correct classification. Figure 11 presents an example of a wrong emotion recognition. The classifier returned Anger when the user's emotion label of the video was Sad. This confusion is predicted since the bottom features inherent to Anger and Sad emotions are identical (Ekman and Friesen, 1975).

Regarding the facial expressions prediction method, in the Figures 8 and 11 we observed that the algorithm correctly estimates eyebrows ”down”, which is confirmed by the original images. The same is detected in the Figure 9 for eyebrows ”up” predictor. Moreover, in the Figure 10, comparing eyebrows of image analyzed and original image, we

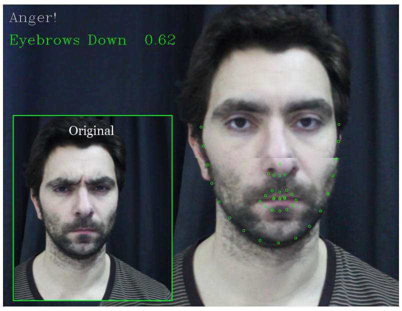

Figure 8: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (green). Check that our emotion and prediction match original image eyebrows movements (green box).

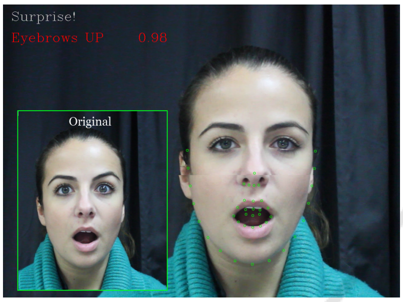

Figure 9: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (red). Check that our emotion and prediction match original image eyebrows movements (green box).

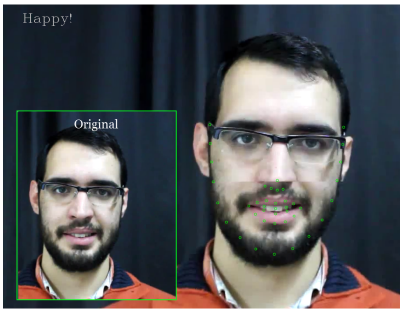

Figure 10: VR Assessing Facial Expressions: Correct Emotion Recognition result (blue) and no Expression Predictor result, since there is not movement. Check original image in green box.

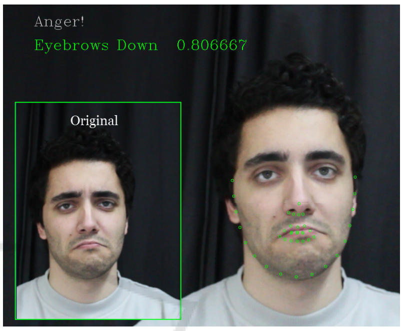

Figure 11: VR Assessing Facial Expressions: Incorrect Emotion Recognition result (blue) and Expression Predictor result. Check original image to see that Expression Predictor is correct (green box).

observe no movement, which traduced in a correct no estimation of movement from both predictors.

5 CONCLUSIONS

This work delivers VR consumer-level methods to achieve the three goals: make MoCap systems compatible to persistent partial occlusions, real-time recognition of universal emotions and real-time prediction of upper face movements using bottom face features tracking. Combining the three methods deployed, we are able to track in real-time facial expressions from non-occluded and occluded facial regions. The development of these methods lead to improvement in the three components of sense of embodiment, i.e. enhances the sense of self-location, agency and body ownership within the VR environments (Kilteni et al., 2012).

Analyzing the results, we conclude that the three goals proposed where achieved. We deliver a method to make MoCap systems able to track bottom face features under generic partial occlusions created by different HMD's. Note, we do not deliver a method that is able to overcome generic and unpredicted facial occlusions, since we require the knowledge of the area occluded. Then, using these facial features, we were able to define methodologies to real-time recognition of four universal emotions (Jack and Jack, 2013) with an accuracy of 69.07% and prediction of facial movements in the occluded regions, i.e. eyebrows "rising" with accuracy of 91.47% and "descending" with an accuracy of 70.63%. The results obtained with the facial expressions prediction method confirmed our method's hypothesis. Therefore, besides bottom features of the face being not enough to describe the six emotions of Ekman and Friesen (Ekman and Friesen, 1975), our predictor of facial expression decode a connection between bottom face and upper face features. As explained in methodology, the combination of both emotion and expressions tracked/predicted make us able to access a wide range of facial expressions enabling us to represent the diversity of faces (McCloud, 1993). This conclusion opens new lines of research to predict more complex movements of the face, even when we are not able to track them using CV algorithms. Furthermore, our methods outputs enable the real-time animation of 3D characters, since we deliver information of facial features combined to emotions, suitable to activate different types of rigs.

Ahead of 3D characters animation, our methods are suitable for emotion-based applications, like affective virtual environments, advertising or emotional gaming.

As future work, we aim to define a transfer algorithm and use movements and emotions estimated to trigger facial animation. Furthermore, we intend to study how the estimation of more facial behaviors information (e.g. forehead and eye movements) and combination of speech data can improve the animation and user embodiment in VR environments.

ACKNOWLEDGEMENTS

This work is supported by Instituto de Telecomunicações (Project Incentivo ref: Projeto Incentivo/EEI/LA0008/2014 and project UID ref: UID/EEA/5008/2013) and University of Porto. The authors would like to thanks Elena Kokkinara from Trinity College Dublin and Pedro Mendes from University of Porto for their support given in the beginning of the project.

REFERENCES

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, 3(2):0–0.

Bombari, D., Schmid, P. C., Schmid Mast, M., Birri, S., Mast, F. W., and Lobmaier, J. S. (2013). Emotion recognition: The role of featural and configural face information. *The Quarterly Journal of Experimental Psychology*, 66(12):2426–2442.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cao, C., Hou, Q., and Zhou, K. (2014). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43.

Cao, C., Weng, Y., Lin, S., and Zhou, K. (2013). 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41.

Eisenbarth, H. and Alpers, G. W. (2011). Happy mouth and sad eyes: scanning emotional facial expressions. *Emotion*, 11(4):860.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto.

Ekman, P. and Friesen, W. V. (1975). Unmasking the face: A guide to recognizing emotions from facial cues.

Fuentes, C. T., Runa, C., Blanco, X. A., Orvalho, V., and Haggard, P. (2013). Does my face fit?: A face image task reveals structure and distortions of facial feature representation. *PloS one*, 8(10):e76805.

Jack, R. E. and Jack, R. E. (2013). Culture and facial expressions of emotion Culture and facial expressions of emotion. *Visual Cognition*, 00(00):1–39.

Kilteni, K., Groten, R., and Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387.

Lang, C., Wachsmuth, S., Hanheide, M., and Wersing, H. (2012). Facial communicative signals. *International Journal of Social Robotics*, 4(3):249–262.

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. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4).

Li, H., Yu, J., Ye, Y., and Bregler, C. (2013). Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4).

Loconsole, C., Runa Miranda, C., Augusto, G., Frisoli, G., and Costa Orvalho, v. (2014). Real-time emotion recognition: a novel method for geometrical facial features extraction. *9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2014)*, 01:378–385.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE.

Magnenat-Thalmann, N., Primeau, E., and Thalmann, D. (1988). Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297.

McCloud, S. (1993). Understanding comics: The invisible art. *Northampton, Mass.*

McCloud, S. (2006). *Making Comics: Storytelling Secrets Of Comics, Manga And Graphic Novels Author: Scott McCloud, Publisher: William Morrow*. William Morrow Paperbacks.

OpenCV (2014).

Pandzic, I. S. and Forchheimer, R. (2003). *MPEG-4 facial animation: the standard, implementation and applications*. Wiley. com.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45.

Parke, F. I. and Waters, K. (1996). *Computer facial animation*, volume 289. AK Peters Wellesley.

Pighin, F. and Lewis, J. (2006). Performance-driven facial animation. In *ACM SIGGRAPH*.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rodriguez, J., Perez, A., and Lozano, J. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575.

Saragih, J. M., Lucey, S., and Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215.

Slater, M. (2014). Grand challenges in virtual environments. *Frontiers in Robotics and AI*, 1:3.

von der Pahlen, J., Jimenez, J., Danvoye, E., Debevec, P., Fyffe, G., and Alexander, O. (2014). Digital ira and beyond: creating real-time photoreal digital actors. In *ACM SIGGRAPH 2014 Courses*, page 1. ACM.

Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011). Realtime performance-based facial animation. *ACM Transactions on Graphics (TOG)*, 30(4):77.