Altering the Conveyed Facial Emotion Through Automatic Reenactment of Video Portraits

Colin Groth
Institut für Computergraphik
TU Braunschweig
groth@cg.cs.tu-bs.de

Jan-Philipp Tauscher
Institut für Computergraphik
TU Braunschweig
tauscher@cg.cs.tu-bs.de

Susana Castillo
Institut für Computergraphik
TU Braunschweig
castillo@cg.cs.tu-bs.de

Marcus Magnor
Institut für Computergraphik
TU Braunschweig
magnor@cg.cs.tu-bs.de

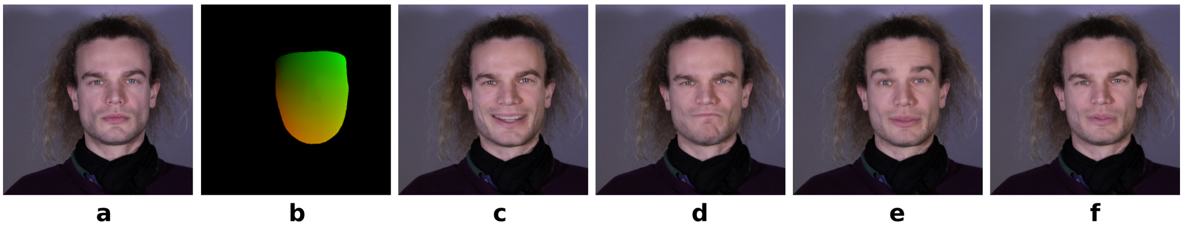

Figure 1: This paper investigates the perception of different emotions in reenacted portrait videos. Using a modified state-of-the-art technique that operates on the *uv* maps (b) of the manipulated meshes, we alter the facial expression of an input video (a) to display happiness (c), disbelief (d), positive surprise (e) and disgust (f).

Please note, that this content is provided by the authors for scientific use and does NOT represent the final version of the paper (final version here).

Abstract

Current facial reenactment techniques are able to generate results with a high level of photorealism and temporal consistency. Although the technical possibilities are rapidly progressing, recent techniques focus on achieving fast, visually plausible results. Further perceptual effects caused by altering the original facial expressivity of the recorded individual are disregarded. By investigating the influence of altered facial movements on the perception of expressions we aim to generate not only physically possible but truly believable reenactments.

When expressing an emotion, individual differences are clear, i.e., two different people can use different facial motions to successfully express the same emotion. However, it is unclear if one can simply map the movements of one subject onto another while preserving the conveyed meaning.

In this paper we perform two experiments using a modified state-of-the-art technique to reenact a video portrait of a person with different expressions gathered from a validated database of motion captured facial expressions to better understand the impact of manipulating the facial expressions via reenactment. Our results indicate that the manipulated faces are not only able to convey the desired emotions but also to preserve the personality of the reenacted individual.

Keywords: facial reenactment, video manipulation, personality perception, facial expressions

1 Introduction

Facial emotions play a key role in communication and are able to dramatically alter the conveyed meaning of a message [1, 2]. People are natural experts in interpreting facial emotions, so mismatches between what is conveyed by the different communication channels are extremely salient for the human audience. Since the biggest part of affective meaning is conveyed non-verbally by facial expressions [3] it is even more important to properly adapt the visual channel when emotions are displayed. Furthermore, in scenarios where there is no real time feedback allowing the speaker to rephrase or clarify possible misunderstandings, – e.g. a recorded video message in opposition to a face-to-face conversation – the only solution to avoid people misinterpreting the displayed emotions can be to re-shoot the video.

In this paper, we focus on a post-processing alternative that exploits the great potential of automatically generated facial reenactments to be an effective and efficient tool to improve quality and comprehensibility of videos, making reshots dispensable in such situations.

We modified an state-of-the-art tool capable of reenacting facial expression in video portraits, by adapting it to be able to use motion capture (MoCap) data as source input instead of videos. To the best of our knowledge, this is the first reenacting technique that uses MoCap data as basis for the manipulation of facial expressions. The presented tool allows to easily reenact facial emotions on user-provided videos with motion captured facial expressions. In the pursuit of the long-term goal of generating novel expressions from existing data, MoCap data provides the genuine advantage of providing precise 3D representations that can be used for multidimensional operations like interpolation between expressions.

Based on the reenacted videos resulting from our technique, we performed two experiments on how reenacted emotions and the arising personalities are perceived. The first experiment focuses on the perception of Recognition, Intensity and Sincerity of the reenacted facial emotions with a focus on the conveyed and intended meaning. The second experiment investigated the influence of manipulated facial movement on the perception of personality. The results demonstrate how reliably specific emotions can be generated and how they will appear to the viewer, especially in comparison with the real videos of the same emotions. We also provide a first view on the perceptual impact on the perceived personality when watching altered videos of an unknown individual.

2 Related Work

Manipulating facial movements in images and videos in a photo-realistic way has become possible with recent advances in computer graphics. These techniques can be used to alter expressions as needed, sometimes even in real-time and usually with a neural network at their core. In particular, Generative Adversarial Networks (GANs) [4] and Auto-Regressive Networks [5] are a frequently used tool to synthesise high quality images. Facial reenactment is usually done by using depth information besides the RGB video input [6, 7], or a parametric model [8].

The first approach to facial reenactment of common RGB videos in real-time was published by Thies et al. [9]. In this approach the authors used a markerless face tracker for face detection without any additional depth information. Although the synthesis of videos is possible at a high quality rate, producing temporally-consistent synthesised videos is a challenging problem that has a long research trajectory. One example for an approach that made the video-to-video synthesis more temporally-consistent is the work of Wang et al. [10]. With their vid2vid framework they were able to produce high-resolution, temporally-consistent video results. Their approach uses a conditional GAN for short-term temporal coherence. Despite the positive characteristics of the vid2vid approach, it is not directly related to faces. A further improvement in facial reenactment in terms of quality was done by Thies et al. with the Deferred Neural Rendering method [11]. This technique allows to use imperfect face meshes and still generate a high quality photo-realistic reenactment. For this, the authors use a neural renderer to interpret the object-specific Neural Textures.

All methods of facial reenactment mentioned above focus mainly on the technical implementation rather than on the correct perception of affective meaning. However, the community is aware of the importance, as it can trigger the uncanny valley effect [12] on the viewer in line with the recent work of Mittal et al. [13]. Thus, a method able to generate emotions that convey the desired meaning while keeping the original look of the person in a photo-realistic manner would have great advantages. Furthermore, given that no human is good (or even capable), to perform all possible facial expressions, as shown in previous research [14, 15], such a tool could help overcome this deficit since the emotions could be subsequently adjusted.

3 Reenacting Technique

Reenactments describe the rendering of images or videos from a new perspective or manipulated aspect, typically done by training a neural network. In this work we investigate the generation of photo-realistic renderings of video portraits with a focus on the representation of the emotional state. The transformations used for the facial reenactment derive from MoCap data.

3.1 Face Tracking

The position of a face in every frame of a video must be known if we aim to manipulate the facial structure of that person. More than that, not only the face position but specific key points like the nose or the corners of the mouth are necessary to effectively manipulate faces in videos.

This work uses face tracking based on facial landmark detection to gather the required information about key point positions in the video portraits. We have chosen this method as it requires relatively few resources and the results are sufficient in most cases.

The tracking that is used to find the face position in the video frames is done by the open-source library Dlib [16]. Dlib was used because it is commonly known and well developed and has a solid trade-off between accuracy and speed.

Since the tracking itself is image-based and, therefore, time independent, it often contains high-frequency noise. To stabilize the face tracking over time an exponential moving average filter is applied on the detected face positions.

The exponential moving average is a good filtering technique for this task because it is an efficient widespread method that, most importantly, can be applied on the data in place.

A three-dimensional mask model of the face is gained by reconstructing it joint-based by a Position Map Regression Network (PRN) as proposed by the work of Feng et al. [17]. This end-to-end method aims to reconstruct the face shape in 3D and jointly predicts the dense alignment of the face. To achieve this goal, a trained encoder-decoder Convolutional Neural Network (CNN) is used to determine the *uv* position maps corresponding to the faces of the input images. Furthermore, an estimation of the head pose relative to the camera position results from the PRN method. The face reconstruction is done at a remarkable speed and is feasible at over 100 FPS [17].

The 3D facial mesh is not only used as an input for the network training but provides the foundation for the facial reenactment process as described in more detail in Section 3.3.

3.2 Motion Capture Data

This work focuses on the use of MoCap data as a basis for facial reenactments instead of using other inputs such as videos. MoCap data has the decisive advantage that it is represented in three dimensions. The benefit of having 3D data is that multiple operations can be done much easier, like interpolating between different emotions to generate movement data for emotions that are not captured at all. Such an interpolation for 2D video inputs would require to generate 3D face representations out of the videos, as expressions are 3D. This dimensional upscaling is still part of current research and not focus of this work.

For the emotion based mesh manipulation, we use the facial expressions MoCap database created by Castillo et al. [18]. This dataset presents several advantages that fit our purposes. It contains 62 different expressions for ten different people with no acting experience. The data is available in two formats: real videos, and their MoCap data synchronous versions. All the emotions were recorded using a method acting protocol [19] which increases their naturalness, and their recognizability has already been validated [15, 18].

For further processing, head motion was separated from the data points so that only the face motion is present in the recordings. This separation was done by specific markers that are not affected by any facial movement but only by head motion. To have the movement separated into head and facial motion is important in the scope of this work, since the final reenactment should only affect the face and manipulate the expressions but not the head position of the person. Such a manipulation would require further processing of the final videos to achieve correct illumination and novel view synthesis where it is needed.

Besides point cloud data, the dataset also features the real videos of the subjects showing the emotions that were recorded by MoCap.

In addition, a second dataset is used containing video portraits of the target person that is to be reenacted. This dataset was captured by us and is composed by a set of twelve different emotions.

3.3 Generation of Novel Expressions

The reenactment of a face from an existing video is done by manipulating or replacing the facial mesh of a target person before delivering it to the renderer that creates the final video. The facial manipulation procedure applied in this work uses the face mesh of the face tracking method described above and the MoCap data as input.

For the further procedure, the two datasets must be comparable because, obviously, the exact same parts of the face shall be manipulated. In order to achieve this comparability we have to know how much each of the 62 MoCap markers influences all of the vertices of the full face mesh when a movement is applied. The weight w is the MoCap marker that influences a particular vertex of the face mesh given by

$$w_{i,j} = (\|X_i - X_j\|^2 * \sum_{k=0}^{N} \frac{1}{\|X_k - X_j\|^2})^{-1}, \tag{1}$$

Here, the 3D points X are indexed by i, j for the MoCap points and mesh vertices, respectively. N equals the total number of MoCap points. The quadratic distance is used as it results in the most accurate emotional representations based on visual investigation.

As both datasets, the face mesh and the MoCap datapoints, are defined in their own spaces, for every frame they are merged into a uniform 3D space before transformation. This unification of the datasets is derived by three predefined points (A, B, C) in each dataset, respectively, that represent the same information in the face, e.g. the tip of the nose. Note, that the goal is to transform all points to a defined normalized position, constructed by the Cartesian space of the face tracking data. Since we are dealing with measured data, the distances between those points can differ slightly in size. Therefore, to be more robust, the transformation is calculated by three vectors $\vec{a}, \vec{b}, \vec{c}$ that derive from A, B, C. They are orthogonal and normalized and further defined by equation 2.

$$\begin{aligned} &\vec{a} = \widehat{BA}, \quad \vec{b} = \widehat{CX}, \quad \vec{c} = \vec{b} \times \vec{a} \\ &\text{with} \quad X = A + \widehat{AB} * (\widehat{AB} \cdot \vec{AC}) \end{aligned} \tag{2}$$

The final transformation of the MoCap point cloud M_p is done by the individual transformations resulting from the calculated vector systems $(\vec{a}, \vec{b}, \vec{c})$ to their origin. For the transformation T_m in the MoCap space and the transformation T_f in the face tracking space the overall transformation $M_{p_{new}}$ is calculated by

$$M_{p_{new}} = M_p * T_m * T_f^{-1}. \tag{3}$$

The face tracking data, as well as the MoCap movements, are normalized for every frame. This implies that the head motion is removed completely from the data, as our current focus is only on facial movements (cf. Sec. 3.2).

The normalization $N(a)$ of a point $a \in F$ from the face tracking data F, computed with the rotation matrix R and translation vector $\vec{t}$, results from the face tracking:

$$N(a) := a * \vec{t} \cdot R \tag{4}$$

The rotation is directly derived from the estimated pose of the tracked person (cf. Sec. 3.1). For translation, a stable point in the face is used.

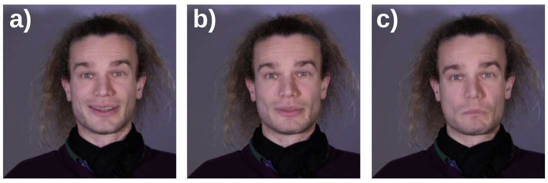

Figure 2: The same emotion (positive surprise) is expressed in different ways: Input video fully reenacted with the expression of the target person (a), based on the MoCap data (c) and based on a 1:1 weight combination of both target person and MoCap data (b). Please note that, even when all of them where correctly recognized in our experiments, the lack of dynamics in this depiction can affect the recognizability of the expression.

The final reenacted face motions result from a weighted combination of the MoCap data and the expressions of the target person extracted from video footage, as follows:

$$M_{final} = w_1 * M_{mocap} + w_2 * M_{target}, \quad (5)$$

with weights $w_1 + w_2 = 1$. M represents the matrix of the n three-dimensional vertices which means that M consists of $n \times 3$ elements. An example of how the different weighting changes the way the emotion is expressed is shown in Fig. 2.

A median filter with a kernel size of five frames is applied on the face mesh movements of the target person before using it in the linear combination of Equation 5. This is to mitigate the impact of movement noise from imperfect tracking and reconstruction. For this task, we use a median filter, which not only is a good fit for the kind of data used but, most importantly, does not dampen the signal.

Finally, the *uv* maps are derived from the reenacted face masks since they are needed for the rendering process described in the following Section 3.4.

3.4 Rendering of Reenacted Videos

We employ Deferred Neural Rendering for novel video synthesis as proposed by Thies et al. [11].

While other techniques require high quality inputs [20, 8], Deferred Neural Rendering makes it possible to produce photo-realistic renderings from imperfect 3D meshes. This is a benefit for our work as the facial landmark-based tracking has some inaccuracies due to measurement errors.

The synthesis of new video frames is done using Neural Textures [11]. Like traditionally learned textures, these feature maps are stored on top of three-dimensional meshes but with additional information.

As inputs for the rendering network, pairs of *uv* maps and their corresponding images are used as illustrated by Fig. 1.

4 Experimental Design

We used a standardized Recognition, Intensity and Sincerity (RIS) framework with forced-choice tasks to examine how reenactment modulates the perception of the expressivity of an individual. This experiment was conducted on-site. Additionally, we investigated the impact of personality on the reenactment by conducting a second experiment using the well-established Five-Factor Model Rating Form (FFMRF) [21].

The on-site experiment considered two different conditions, both with a different group of participants. In the first condition (C1) participants were only exposed to the real videos of both actors (the target actor and die MoCap actor) showing the chosen emotions. For the second condition (C2) the reenacted videos of the same chosen emotions were used as stimuli.

The online experiment had five conditions each with a different type of reenactment and covers more than 700 participants.

Stimuli Two datasets were used for the generation of the stimuli displayed in both experiments. The first dataset was the MoCap dataset with the corresponding real videos as described in Sec. 3.2 while the second consisted of video portraits of the target person. For the experiments four representative emotions from both datasets were selected: *happiness, positive surprise, disbelief* and *disgust*. Other more common emotions, such as anger, were not selected if they had no strong representation in the Mo-

Cap data or were mainly expressed through head motion. Please note that two of the selected expressions are positive emotions while the other two are negative emotions, and none of them necessarily requires head motion to be correctly recognized [14, 22].

Additionally, a neutral video of the target person was used as basis for the reenactments. This video shows the person with neutral or dimmed expressions while talking. It contained very little head motion to prevent interferences with the reenacted facial expressions and consequential misinterpretations of the participants in the experiments.

In the on-site experiment, three different types of reenactments were used for the second condition (C2), while the online experiment contained two more types of reenactment (5 in total). The reenactments differed in their representation of the weighted combination of w_1 and w_2 (cf. equ. 5):

1. $w_1 = 0$ and $w_2 = 1$ ($JPT_{re}; R_1$). These expressions are only based on the facial movement of the target person. They are to show how stable the render method is and how much noise is introduce by using MoCap.

2. $w_1 = 1$ and $w_2 = 0$ ($MOCAP_{re}; R_5$). The expressions in these videos are fully based on the MoCap data and demonstrate how good emotions are recognised when they are transferred from MoCap.

3. $w_1 = 0.5$ and $w_2 = 0.5$ ($Mix_{re}; R_3$). These expressions consist of the movement of the target person and the MoCap data equally and will allow a comparison of the two sources.

4. $w_1 = 0.75$ and $w_2 = 0.25$ (R_4, online only). These expressions are in between the full movement transfer and the equal movement division and will allow further analysis.

5. $w_1 = 0.25$ and $w_2 = 0.75$ (R_2, online only). These expressions are in between the full geometry representation and the equal movement division and will allow further analysis.

All reenacted videos as well as the real videos were displayed from and to neutral expression when showing an emotion. The length of the videos varied between two and five seconds.

Apparatus All trials of the on-site experiment were conducted using a 24-inch screen (1920 x 1080 px, 60 Hz) while the participants sat alone in a closed room to prevent external distractions and ensure that the participants choose their answers freely. The online experiment was implemented using the LimeSurvey software [23] and distributed via Amazon Mechanical Turk [24].

The videos for both the training and all reenactments were post-processed to 720 x 720 px at 50 fps.

Participants For the real videos condition (C1) a total number of 21 participants (11 females, age 18–32, mean 24.1, SD 3.05) and for the reenacted videos condition (C2) a total number of 22 participants (10 females, 19–57, mean 26.45, SD 11.01) attended.

The majority of participants reported no former experience in computer graphics.

The total number of participants for the online experiment after sanity checking was 710 (out of 829). Participants were assigned randomly to the conditions with at least 127 sessions each.

All participants of the on-site experiment were compensated for their time with 10€ or equal internal university's course credit.

Procedure The conditions of the on-site experiment differ in types of videos that were shown to the participants (cf. Section 4) but followed the same procedure, controlled by Psychophysics Toolbox Version 3.0.11 (PTB-3) [25].

After fulfilling an informed consent and collecting demographic data, each participant was placed in the experiment room and received a task introduction to judge a video by the first impression.

The concepts of RIS were explained on screen at the beginning of the experiment. During a trial, participants watched the videos of the four emotions (*happiness*, *positive surprise*, *disbelief* and *disgust*) for every actor, as mentioned in Section 4. The presentation of the stimuli followed a blockwise design by type of reenactment. Within a type, the order of presentation of the emotions was fully randomized, with each participant being assigned a different order. For all videos, no repetitions were possible to force the viewers to decide by their first impression.

For every video, the participants were asked to answer three multiple-choice questions: ”Which emotion is expressed?”, ”How intense is the emotion expressed?”, ”How sincere is the emotion expressed?”. While the first question was categorical (forced choice), the other two were asked to be rated on a seven-point Likert-scale going from ”extremely low” to ”extremely high”. After finishing all tasks the results were checked for completeness and the participants were compensated.

The procedure in the online experiment followed the one from the on-site experiment with the difference that participants only watched the four emotions in the reenacted videos and fulfilled the FFMRF afterwards. The participants were instructed to fulfill the FFMRF on a 7-point rating scale for every personality trait.

5 Results and Discussion

Overall, the results of the experiments indicate that it is possible to manipulate conveyed emotion through facial reenactment while the original personality is preserved. This is true even for combined sources of movement.

RIS Fig. 3 illustrates the results of the on-site experiment for both conditions. The results of the three types of reenactments are shown side by side with the two ground truth conditions for better visualization.

For the analysis of the experimental results, a two-way ANOVA per averaged dimension with type of reenactment and emotion as within-participant factors was done for each of the two conditions.

The condition for the reenacted videos (C2) showed a significant effect on the emotion for the **Recognition**, $F(3, 336) = 16.5, p < 0.001$. In comparison, for the condition with the real videos (C1) there was no significant effect on the emotion. All these results were independent from the technique used. From these results it can be concluded that the type of an emotion displayed in a reenacted video matters in terms of recognisability. In this experiment, all emotions from the real videos were well recognised for both actors with average recognition rates over 80% indicating that participants had no problems identifying these emotions. The results from the reenactments of C2 show comparable ratings. All emotions were well recognised except for *disgust*. The recognition rates of *disgust* are only around chance level for all reenactment styles suggesting that participants were not able to detect this emotion. In particular, positive emotions were better identified than negative ones based on the recognition rates also being in line with previous research findings [26, 15]. The rates of *positive surprise* even surpassed the results of the ground truth results of the target actor (JPT_{rv}).

A closer look showed that in those scenarios where the judgements of the participants were not accurate regarding recognition, it took the participants about ten times longer to vote, indicating their insecurity about their judgements. *Disgust* was not generally mistaken for another emotion, but the rates are for the most part evenly distributed among the possible answers. This results indicate that the participants generally did not mistake this emotion for another, but guessed it instead. Only for the reenactments with movements of the MoCap actor ($MOCAP_{re}$ and Mix_{re}) a trend of the wrong answers toward *disbelief* occurred (50.0% of the ratings) that supports the confusion between the two negative emotions, which also appeared in the classification of *disbelief* itself (around 40% mistaken for *disgust*). A possible reason for the bad recognition of *disgust* might that it involves more than only facial motion and needs additional information to be recognised. As other motions were not reenacted in the videos the required information for recognition of the emotion may have been lost.

The experiment also demonstrates that the **Intensity** of the conveyed emotion can be altered. The results of the experiment show a significant effect for the reenactments (C2) for both style and emotion, all $F's > 26.25$, all $p's < 0.001$. For the real videos of C1 an effect was also present for emotions ($F(3, 160) = 3.16$, $p < 0.02$). In contrast to recognition, the reenactment technique has an influence here. The intensities of the emotions from the real videos were all ranked above the average ”neutral” value (score 4) on the Likert scale.

Comparing this to the C2 ratings, it can be seen that the assessments made for *positive surprise* and *happiness* are almost identical to their

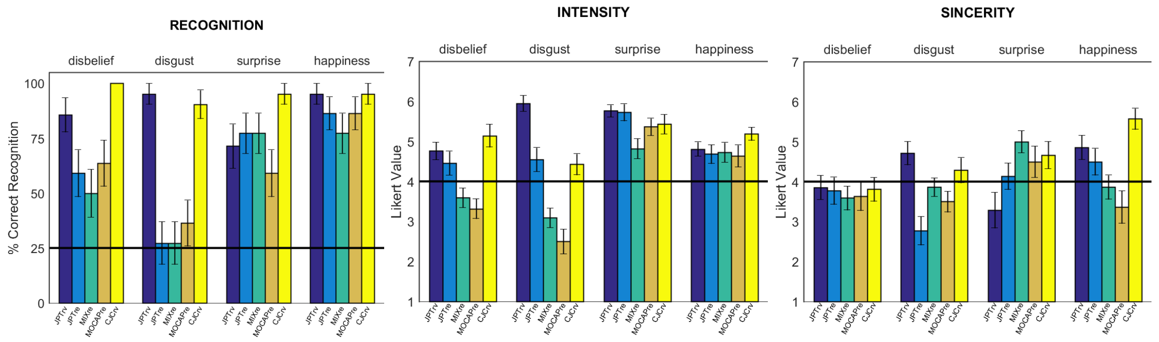

Figure 3: Ratings for the reenacted videos (C2: $JPT_{re}, MIX_{re}, MOCAP_{re}$) anchored by the results from the original videos (C1: JPT_{rv}, CJC_{rv}). The three graphs display in reading order: the recognition rates per expression and type of video; the ratings for their perceived intensity; and the ratings for perceived sincerity. The error bars represent the standard error of the mean (s.e.m.), the chance line is drawn in black.

comparatives. This indicates that positive emotions can be transferred effectively through reenactment. For the emotion *disgust* the ratings are quite low. This is not surprising considering that this emotion was not recognised reliably. The ratings for *disbelief* show a clear distinction between the different styles. For the reenactment done by the full input of the target actor (scale JPT_{re}) the intensity ratings are comparable to the ones of the real videos of the same person. In contrast, the reenactments fully resulting from the MoCap data differ from its comparative rating by the real videos. This deviation indicates that the MoCap data of this expression alters the intensity of the real emotion. This finding is reinforced by the fact that the rating of the intensity of the equal weight combination (namely Mix_{re}) of both styles lies between both. This distribution between the different movement weightings is apparent for almost all emotions and shows the clear possibility that the intensity can be altered by the amount of MoCap movements used.

For the **Sincerity** of the emotions, the experiment indicates that a manipulation by reenacting a video is not possible. For both conditions, the experiment results demonstrate a significant effect of emotion for the sincerity of expression, $F(3, 160) = 7.52, p < 0.001$ (C1) and $F(3, 336) = 7.39, p < 0.001$ (C2). However, no effect of style could be detected. The analytical results clearly reveal that the sincerity of an emotion is not affected by the performed reenactment. This means that the sincerity of facial movements does not get lost when manipulating a video by reenactment. This result presents a positive perspective for the use of reenactments to create believable result videos. In other words, an emotional expression may not be made more authentic by reenactment. Nevertheless, it must be noted that this result depends heavily on the emotion chosen. For *disbelief* and *positive surprise* a well balanced distribution can be seen which is aligned to the values of the real videos. For the other two dimensions, however, the results are not aligned with the real videos, even though they are evenly distributed over the different movement proportions of the reenactments. This divergent behaviour shows that a deliberate manipulation of the sincerity through reenactments is not possible.

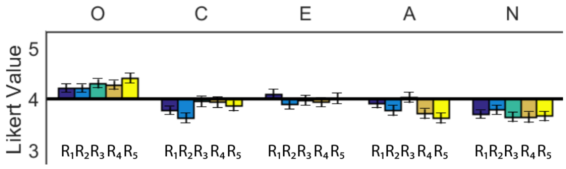

Figure 4: Results of the online experiment per dimension and reenactment type. The s.e.m. is represented by error bars. Note, that only a part of the rating scale is shown, with *4* being ”neutral”.

Personality The online experiment had five conditions, one per type of reenactment ($R_i, i \in (1..5)$) as described in Sec. 4.

This experiment introduced two more conditions compared to the on-site experiment to get a clear gradation of the movement proportions and as a result a clear declaration about the change of personality. The influence of the facial geometry (e.g. look of that person) and movement of the face for the perception of personality now results from the change of the personality ratings along one dimension.

A one-way ANOVA per personality dimension with the facial expression as a within-participant factor and the type of reenactment as a between-participants factor showed if the five personality factors are affected when the source of movement is changed.

For the personality trait **Openness (O)** the analysis states that the personality is preserved ($F = 0.805$, $p = 0.5219$). The **Conscientiousness (C)** of a person is not affected by the manipulation of the movement based on the results of the analysis ($F = 1.805$, $p = 0.1259$). The perceived **Extroversion (E)** of a person seems to rely on the facial geometry when expressing emotions ($F = 0.952$, $p = 0.4330$). For **Agreeableness (A)** the ANOVA indicates a significance for the change of movement ($F = 3.210$, $p = 0.0125$). The **Neuroticism (N)** of a person in a reenacted video is not significantly affected by the source of facial movement ($F = 0.374$, $p = 0.8267$).

Since *Agreeableness* was the only dimension showing a significant change in movement, a *Tukey Honest Significance Difference Test (HSD)* was conducted. The results of the Tukey HSD state that a significance in the personality dimension of *Agreeableness* is only present for the comparison of the conditions with reenactments R_3 and R_5. In every other combination no significant effect of movement occurs. From this result it can be concluded that overall the movement of the face does not play a decisive part for this personality dimension either. The results of this study demonstrate that reenactments are not only able to alter the conveyed meaning of an expression but also preserve the personality of that person even when the facial movement fully comes from a different person.

6 Conclusion

In this paper, we presented insights about the meaning of facial emotions and the effect of personality when the movement components are altered. The input videos were reenacted using our tool which combines different facial movements including MoCap data to create the desired expressions.

The experimental results indicate that reenacted videos are able to manipulate the conveyed facial emotions by generating expressions that are recognised correctly. From the experiment follows that the movements of one person can be mapped onto another and still preserve the conveyed meaning.

We also showed that the intensity of emotions can be influenced using video reenactment. With this perspective, reenactments can be designed in order to ensure that the intensity of the conveyed emotion is appropriate and matches the desired level. This poses not only a great opportunity but also an interesting insight into the human perception and processing of artificially generated emotions. The sincerity of the source emotion is preserved in the facial reenactment and cannot be deliberately changed. This prevents artificially generated expressions from appearing as faked or being used as forgery.

Furthermore, we found empirical evidence that the perceived personality of an individual is preserved when altering the facial expressions of that person with the movements of an actor with a different personality. The analysis of the five dimensions of personality clearly stated a predominance of geometrical information (i.e., the facial features of the target subject) over facial movement as the key factor in the perception of personalities. This preservation of personality is especially important if we are familiar with a person as it can otherwise trigger the uncanny valley effect.

In the future, we will extend our work to interpolate between different emotions to create novel expressions for which no data is available. The realization of this extension is now much easier to accomplish using our validated tool that already works with 3D MoCap data. Furthermore, we would like to including more actors in future experiments and study if the discovered effects chance when different actors are reenacted.

Although the future possibilities are not exhausted, we believe this work provides significant insights into the perception of emotions and personality and their possible manipulation through facial reenactment.

Acknowledgments

The authors gratefully acknowledge funding by the German Science Foundation (DFG MA2555/15-1 “Immersive Digital Reality”).

References

[1] E. Krahmer, Z. Ruttkay, M. Swerts, *et al.* Pitch, eyebrows and the perception of focus. In *SSP*, 2002.

[2] W. Condon and W. Ogston. Sound film analysis of normal and pathological behaviour patterns. *J Nerv Ment Dis*, 143:338–347, 1966.

[3] A. Mehrabian and S. Ferris. Inference of attitudes from nonverbal communication in two channels. *J Consult Clin Psychol*, 31(3):248–252, 1967.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.* Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

[5] A. Van den Oord, N. Kalchbrenner, L. Espeholt, *et al.* Conditional Image Generation with PixelCNN Decoders. In *NeurIPS*, pages 4790–4798, 2016.

[6] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, *et al.* Being John Malkovich. In *ECCV*, pages 341–353, 2010.

[7] T. Weise, S. Bouaziz, H. Li, *et al.* Realtime performance-based facial animation. *ACM ToG*, 30(4):1–10, 2011.

[8] K. Dale, K. Sunkavalli, M. Johnson, *et al.* Video face replacement. *ACM ToG*, 30(6):1–10, dec 2011.

[9] J. Thies, M. Zollhöfer, M. Stamminger, *et al.* Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*, pages 2387–2395, 2016.

[10] T. Wang, M. Liu, J. Zhu , *et al.* Video-to-video synthesis. In *NeurIPS*, 2018.

[11] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM ToG*, 38(4):66:1–66:12, July 2019.

[12] M. Mori, K. MacDorman, and N. Kageki. The uncanny valley. *IEEE Robot Autom Mag*, 19(2):98–100, 2012.

[13] T. Mittal, U. Bhattacharya, R. Chandra, *et al.* Emotions don’t lie: A deepfake detection method using audio-visual affective cues, 2020.

[14] D. Cunningham, M. Kleiner, C. Wallraven, *et al.* Manipulating video sequences to determine the components of conversational facial expressions. *ACM TAP*, 2(3):251–269, 2005.

[15] S. Castillo, C. Wallraven, and D. Cunningham. The semantic space for facial communication. *CAVW*, 25(3-4):223–231, 2014.

[16] D. King. Dlib-ml: A machine learning toolkit. *J Mach Learn Res*, 10:1755–1758, 2009.

[17] Y. Feng, F. Wu, X. Shao, *et al.* Joint 3D face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[18] S. Castillo, K. Legde, and D. Cunningham. The semantic space for motion-captured facial expressions. *CAVW*, 29(3-4):e1823, 2018. e1823 cav.1823.

[19] K. Kaulard, D. Cunningham, H. Bülthoff, *et al.* The MPI Facial Expression Database — A Validated Database of Emotional and Conversational Facial Expressions. *PLoS ONE*, 7(3):e32321, 03 2012.

[20] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.

[21] S. Mullins-Sweatt, J. Jamerson, D. Samuel, *et al.* Psychometric properties of an abbreviated instrument of the five-factor model. *Assessment*, 13(2):119–137, 2006.

[22] M. Nusseck, D. Cunningham, C. Wallraven, *et al.* The Contribution of Different Facial Regions to the Recognition of Conversational Expressions. *J Vis*, 8(8):1,1–23, 06 2008.

[23] LimeSurvey. https://www.limesurvey.org/. Accessed: 2020-05-12.

[24] Amazon Mechanical Turk. https://www.mturk.com/. Accessed: 2020-05-12.

[25] D. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

[26] D. Cunningham and C. Wallraven. The interaction between motion and form in expression recognition. In *APGV*, pages 41–44, 2009.