Automatic Replication of Teleoperator Head Movements and Facial Expressions on a Humanoid Robot

Jan Ondras[1], Oya Celiktutan[2], Evangelos Sariyanidi[3] and Hatice Gunes[1]

***Abstract*— Robotic telepresence aims to create a physical presence for a remotely located human (teleoperator) by reproducing their verbal and nonverbal behaviours (e.g. speech, gestures, facial expressions) on a robotic platform. In this work, we propose a novel teleoperation system that combines the replication of facial expressions of emotions (neutral, disgust, happiness, and surprise) and head movements on the fly on the humanoid robot Nao. Robots' expression of emotions is constrained by their physical and behavioural capabilities. As the Nao robot has a static face, we use the LEDs located around its eyes to reproduce the teleoperator expressions of emotions. Using a web camera, we computationally detect the facial action units and measure the head pose of the operator. The emotion to be replicated is inferred from the detected action units by a neural network. Simultaneously, the measured head motion is smoothed and bounded to the robot's physical limits by applying a constrained-state Kalman filter. In order to evaluate the proposed system, we conducted a user study by asking 28 participants to use the replication system by displaying facial expressions and head movements while being recorded by a web camera. Subsequently, 18 external observers viewed the recorded clips via an online survey and assessed the quality of the robot's replication of the participants' behaviours. Our results show that the proposed teleoperation system can successfully communicate emotions and head movements, resulting in a high agreement among the external observers ($ICC_E = 0.91$, $ICC_{HP} = 0.72$).**

I. INTRODUCTION

Robotic telepresence offers a convenient substitute for face-to-face communication as it provides physical embodiment at a remote place and allows the teleoperator to express non-verbal cues such as head movements, hand gestures, and facial expressions along with audio cues to another person via a robot. Facial and head cues carry significant information with regard to an individual's social signals including emotions, personality and intentions, which are key in enabling effective communication. Remotely located team members are less included in co-operative activities than co-located team members [1], and have fewer conversational turns and speaking time in group conversations [2]. These shortcomings can be mitigated by using robots for telepresence in various settings including remote education [3] and elderly care [4].

Many studies have addressed the automatic recognition of facial expressions and the tracking of head pose in the context of human behaviour analysis and human-computer interaction [5]. However, (i) little work has been done to apply such computational methods to robot teleoperation; and (ii) only a few studies have focused on the investigation of how teleoperators are perceived from the perspective of their interlocutors [6], [7], [8].

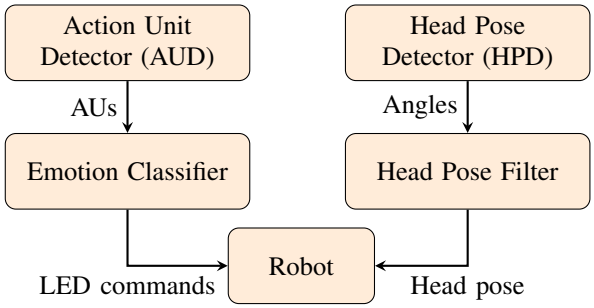

Fig. 1. Main components of the proposed replication system.

In this paper, we propose a novel replication system that reproduces facial expressions and head movements of the teleoperator on a robot avatar in an automatic manner. As shown in Fig. 1, our system is composed of two phases: 1) Analysis phase – recognition of facial expressions and estimation of head movements from a video stream; and 2) Synthesis phase – imitation of the recognised emotions and the estimated head movements via the robot avatar. We evaluated the performance of the replication system both quantitatively and qualitatively by conducting a number of computational experiments and a user study. We first asked a set of participants to use the replication system by displaying facial expressions and head movements while being recorded by a web camera. We then asked external observers to view the recorded clips using an online survey, and assess the quality of the robot's replication of the participants' non-verbal head and face behaviours. Our results show that the proposed teleoperation system can successfully communicate emotions and head movements, resulting in a high rater agreement among the external observers ($ICC_E = 0.91$, $ICC_{HP} = 0.72$).

II. RELATED WORK

Expression of emotions via robots has been a popular research area. In one prominent work, Chevalier et al. [9] compared two robotic platforms with a virtual agent and a human for animating four emotions (anger, happiness, fear and sadness) through facial and bodily cues. These robotic platforms were Zeno and Nao, however, only bodily cues were considered for the Nao robot. They found that facial

[1]J. Ondras and H. Gunes are with the Computer Laboratory, University of Cambridge, UK. jo356@cam.ac.uk; Hatice.Gunes@cl.cam.ac.uk

[2] O. Celiktutan is with the Personal Robotics Lab, Imperial College London, UK. This work was completed while she was with the University of Cambridge. o.celiktutan-dikici@imperial.ac.uk

[3]E. Sariyanidi is with the Centre for Autism Research, Philadelphia, US. sariyanidi@gmail.com

cues play a more important role in conveying emotions, and sadness is the easiest emotion to recognise. Johnson et al. [10] demonstrated that the Nao robot can satisfactorily imitate human emotions through facial cues. This was done by altering the colour, intensity, sharpness and orientation in the Nao robot's eyes. Song et al. [11] studied the effects of three interaction modalities (colour, sound, and vibration) on human perception of emotions. They found that the colour modality is the most important channel for communicating affect and using all modalities simultaneously improves the results. Therefore, we used the colour modality for conveying emotions. Compared to our work, the above-mentioned studies displayed the emotions in various ways and investigated their perception, but they did not deal with the real-time replication of expressions of a human operator.

Despite the importance of facial cues in interaction, previous teleoperation studies have mostly focused on portraying bodily cues on a robot avatar. For example, Bremner and Leonards [12] demonstrated the utility of iconic gestures using a real-time skeleton algorithm using a Kinect depth sensor. In the TERESA project [13], Shiarlis et al. developed a teleoperation system to allow elderly people to participate in social events remotely. The developed robot was able to semi-autonomously navigate among groups, maintain face-to-face contact during conversations, and display appropriate body poses. Agarwal [14] developed a real-time system for the imitation of human head movements. They recorded the teleoperator's head motion by a Microsoft Kinect sensor, and processed these recordings in three steps: (i) low-pass pre-filtering; (ii) neural network-based head pose mapping (model between the Kinect and OptiTrack - ground truths from more accurate sensor); and (iii) constrained-state Kalman filtering. This work constitutes a pioneering effort for head pose replication.

In this paper, we follow a similar approach to the head pose replication of [14], but with the following contributions: (i) we use a regular RGB web camera, and an automatic Head Pose Detector (HPD) to track the head movements; (ii) we propose a novel emotion replication method based on automatic action unit detection; (iii) we collect an in-house dataset for system evaluation; and (iv) we present extensive experimental results for each component, namely action unit detection, emotion classification and head pose filtering, as well as a user study assessing the performance of the teleoperation system by its potential users.

III. METHODOLOGY

As illustrated in Fig. 1, the proposed replication system consists of five main components that can be grouped under two phases, namely, analysis and synthesis. In the analysis phase, Action Unit Detector (AUD) and Head Pose Detector (HPD) simultaneously process incoming video stream frame by frame. The emotion classifier infers the displayed expressions of emotion (e.g. happiness, surprise, etc.) from the detected action units (e.g. brow lowerer, lip corner puller), and head pose filter mitigates the effect of noisy head pose estimations. Both of these components send commands directly to the robot, and the synthesis phase enables the robot to exhibit the target behaviour.

A. Emotion Classification

1) Action Unit Detector (AUD): Human emotions are known to be expressed by facial muscle movements. Ekman et al. proposed the Facial Action Coding System (FACS) that encodes these facial muscle movements in terms of the activation of a set of predefined action units (AUs) [15]. This system forms the basis of a significant number of automatic emotion recognition methods. For example, in a simple rule-based method, happiness can be represented as a combination of AU6 (cheek raiser) and AU12 (lip corner puller) [16].

There has been a significant body of work in the area of automatic AU detection. Sariyanidi et al. [5] highlighted the importance of two practices: (i) combining shape and appearance features, which yields better performance because these feature types carry complementary information, and (ii) using differential features, i.e. features that describe information with respect to the neutral face. The main advantage of the differential features is to place higher emphasis on the facial action by reducing person-specific appearance cues.

Following these insights, in this paper, we used four types of features, namely, shape, appearance, differential-appearance and differential-shape features. Shape features were obtained by concatenating the vertical and horizontal coordinates of the facial landmarks that were estimated using the method of [17]. Differential-shape features were computed by subtracting the shape representation of a given facial image from the shape representation that was computed from a facial image of the same subject, with a neutral expression. Appearance features were extracted using the Quantized Local Zernike Moments (QLZM) method in a part-based manner [18]. Using the estimated facial landmarks, we first cropped three square patches that contained the left eye, right eye and mouth, and then computed the QLZM histograms from each patch. Finally, we computed differential appearance features using the Gabor motion energy [19], where we adopted a part-based representation similarly to the appearance features. We used Gabor energy to describe the motion between a given facial image of a subject and the subject's neutral facial image. An advantage of the Gabor representation, compared to using simpler representations (e.g. difference between neutral and expressive image), is its robustness to illumination variations. For detection, we trained four binary SVM classifiers, each in conjunction with one of the abovementioned feature types, per AU. The final AU detection decision was obtained by fusing the outputs of the four individual classifiers. Specifically, we adopted the *consensus fusion* approach, where an AU was detected based on the condition that all four classifiers were in full agreement. AUD delivered a binary sequence of the detected AUs in a frame-by frame manner. These outputs were then accumulated over the last W frames, and were fed into the emotion classifier for real-time inference.

2) Emotion Classifier: For emotion classification, we trained a neural network with one hidden layer. Specifically,

we used a multilayer perceptron model with rectified linear unit (ReLU) as an activation function for the hidden layer and softmax for the output node. The cross-entropy loss function was minimized by solver L-BFGS, which uses an approximation of inverse Hessian matrix to steer the search.

In order to generate input feature vectors for the neural network, we summarized the detected AUs over the last W frames using two different strategies: 1) **AVG**: the last W frames were averaged per AU, generating a feature vector x of length N; and 2) **CONCAT**: the last W frames were concatenated, creating a feature vector x of length $N \times W$, where N was the number of AUs.

B. Head Pose Filtering

Head Pose Detector (HPD) was based on the method introduced in [17]. However, raw head pose estimations from HPD typically contain significant noise, which results in jerky robot head movements. Therefore, we further applied a head pose filter in order to achieve a smooth head pose trajectory, in terms of angular position, velocity and acceleration, that can be displayed on the robot by taking into account the robot's movement constraints. This was addressed by constrained-state Kalman filter with a minimum jerk model, similarly to Agarwal et al. [14].

The state vector at a time instant t can be defined as

$$x_t = \begin{bmatrix} \theta_t & \omega_t & \alpha_t \end{bmatrix}^T$$

where θ_t is the angle, ω_t is the angular velocity, and α_t is the angular acceleration. It has been shown that the voluntary human movements obey a minimized jerk trajectory, which is also the smoothest trajectory [20], [21]. Therefore, we assume the minimum jerk model as a transition model. Let's consider the system given by

$$\begin{aligned} x_{t+1} &= F_t x_t + w_t \\ y_t &= H x_t + v_t. \end{aligned}$$

where

$$F_t = \begin{bmatrix} 1 & \Delta T_t & \frac{\Delta T_t^2}{2} \\ 0 & 1 & \Delta T_t \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

is the transition model, ΔT_t is the time step, y_t is the measurement, H is the measurement model (in our case $H = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, since we measure angle θ_t only), $w_t \sim \mathcal{N}(0, Q)$ is the Gaussian process noise (at acceleration level only), and $v_t \sim \mathcal{N}(0, R)$ is the Gaussian measurement noise.

The Kalman Filter is described using the following:

1) Time update (prediction)

$$\begin{aligned} P_t^- &= F_t P_{t-1}^+ F_t^T + Q \\ \hat{x}_t^- &= F_t \hat{x}_{t-1}^+ \end{aligned}$$

2) Measurement update (correction)

$$\begin{aligned} K_t &= P_t^- H^T \left(H P_t^- H^T + R \right)^{-1} \\ P_t^+ &= \left(I - K_t H \right) P_t^- \\ \hat{x}_t^+ &= \hat{x}_t^- + K_t \left(y_t - H \hat{x}_t^- \right) \end{aligned} \tag{2}$$

where $\hat{x}_t^-$ is the a priori estimate of state x_t given measurements up to time $t-1$ inclusive, $\hat{x}_t^+$ is the a posteriori estimate of x_t given measurements up to time t inclusive, P_t^- is the covariance matrix of the a priori estimation error $(x_t - \hat{x}_t^-)$, P_t^+ is the covariance matrix of the a posteriori estimation error $(x_t - \hat{x}_t^+)$, and K_t is the Kalman gain.

The output $\hat{x}_t^+$ of the original filter given by (2) might be truncated due to the robot's movement constraints, resulting in discontinuities in the robot's head movements. To overcome this problem, we use constrained-state Kalman filter, and project the unconstrained estimate $\hat{x}_t^+$ onto the constraint surface [22]. The constrained state estimate $\tilde{x}_t^+$ can then be formulated as an optimization problem,

$$\tilde{x}_t^+ = \arg\min_x \left(x - \hat{x}_t^+ \right)^T U \left(x - \hat{x}_t^+ \right) \tag{3}$$

subject to inequality constraints $\begin{bmatrix} \theta_{min} & \omega_{min} & \alpha_{min} \end{bmatrix}^T \leq x$ and $x \leq \begin{bmatrix} \theta_{max} & \omega_{max} & \alpha_{max} \end{bmatrix}^T$. Here, θ_{min}, ω_{min}, α_{min}, θ_{max}, ω_{max}, α_{max} define the known fixed bounds on state. Setting $U = \left(P_t^+ \right)^{-1}$ guarantees the maximum probability state estimates (i.e. minimum variance filter) subject to the state constraints.

C. Robotic Platform

For the robotic platform, we used the humanoid robot Nao developed by Aldebaran Robotics [23] with the technical details of NaoQi version 2.1, head version 4.0 and body version 25. The Nao robot has a static face, and cannot display facial muscle movements. However, in [10], Johnson et al. demonstrated that the Nao robot can use LED colours and patterns to imitate emotions. Inspired by this study, we mapped each emotion onto a different colour code that was displayed on Nao using its eyes' LEDs – i.e. neutral emotion with no colour, disgust with the colour red, happiness with the colour blue, and surprise with the colour green.

IV. EXPERIMENTS AND RESULTS

We evaluated the proposed replication system both quantitatively and qualitatively. In this section, we first present our experimental results with regard to the performance of AU detector, emotion classifier and head pose filter on the publicly available databases. We then provide analysis and results from the user study conducted.

A. Action Unit Detection

In this work, we focused on the following seven AUs, namely, inner brow raiser (AU1), outer brow raiser (AU2), brow lowerer (AU4), cheek raiser (AU6), lip corner puller (AU12), parted lips (AU25), and jaw drop (AU26), as only these 7 AUs (out of a maximum 12 detectable by AUD) are found to be ever active on the CK+ dataset. For each AU, we trained an SVM classifier using the one-vs-all approach using linear c-SVM [24] and fixing the c parameter to $c = 10^{-3}$. We evaluated the performance of the trained AU detector using the MMI Facial Expression database [25], one of the most widely used benchmark datasets in the field. In Tab. I, we reported the 5-fold subject-independent cross validation

results for the four individual features and their combination with consensus fusion in terms of the two-alternative forced choice (2AFC) metric [26]. 2AFC metric can be defined as as the area A underneath the receiver-operator characteristic (ROC) curve, as well as an upper bound for the uncertainty of the A statistic for n_p positive and n_n negative samples, $s = \sqrt{(A(1-A)/min\{n_p, n_n\})}$. Looking at Table I, the best performing individual feature was the differential-appearance feature, and the consensus fusion further improved the 2AFC score as compared to using differential-appearance feature alone for the detection of four AUs - AU1, AU6, AU12, and AU26.

TABLE I

AU DETECTION PERFORMANCE IN TERMS OF 2AFC SCORE (BOLD INDICATES THE HIGHEST SCORE). (δ: DIFFERENTIAL)

	AU1	AU2	AU4	AU6	AU12	AU25	AU26
Shape	0.74	0.53	0.67	0.61	0.79	0.73	0.53
Appear.	0.74	0.73	0.65	0.78	0.82	0.78	0.67
δ-Shape	0.78	0.67	0.71	0.74	0.78	0.82	0.64
δ-Appear.	0.90	**0.92**	**0.87**	0.82	0.92	**0.89**	0.78
Fusion	**0.91**	0.89	0.78	**0.87**	**0.93**	0.86	**0.79**

B. Emotion Classification

We trained the emotion classifier using 294 image sequences from the publicly available CK+ database [27], another widely used benchmark dataset in the field. The CK+ database contains image sequences that start with a neutral face and end at the peak intensity of a target emotion enabling the calibration of the AU detector using the first frame. Each sequence is labeled with one target emotion from the seven basic emotions, and has a length of 6–65 frames with the mean of 16 frames.

We implemented the neural network using Python machine learning library *sci-kit learn* [28] and used a stratified 5-fold cross-validation to tune the following hyperparameters: hidden layer size h, window size W, and regularization parameter α. We repeated the same strategy for both input feature types AVG and CONCAT.

Our preliminary results showed that AVG representation yielded a higher recognition rate as compared to CONCAT representation on the validation set (i.e. $AVG = 83\%$ and $CONCAT = 80\%$). For the four emotions in question, the emotion classifier provided us with the following recognition accuracy: neutral 59%, disgust 75%, happiness 81% and surprise 90%.

For the real-time version of the proposed system we trained the neural network using the AVG representation (with optimal hyperparameters set during cross-validation as $hls = 3$, $W = 5$ and $\alpha = 10^{-3}$) and the whole database for recognising four emotions.

We built an in-house dataset to test the trained model on a set of unseen samples. We recruited a total of 28 participants in two separate groups, and asked them to perform the four emotions twice in front of a web camera: Group 1 – 16 participants, starting with a neutral face followed by the display of the target emotion ($16 \times 4 \times 2 = 128$ test clips in total); and Group 2 – 12 participants displaying only the target emotion ($12 \times 4 \times 2 = 96$ test clips in total). Representative examples are shown in Fig. 2.

In total, we collected 224 test clips, each had a duration of approximately 15 secs. The emotion classifier scanned each test clip over a sliding window of W, and delivered the recognised emotion at each time step. The final decision for the whole clip was then made by majority voting of the outputs from all temporal windows.

Fig. 2. Representative examples from the recorded dataset: *disgust* (top) and *surprise* (bottom).

We evaluated the emotion classifier using two approaches: 1) AUD was *not* provided with a calibration clip, and used the first frame as a reference for calibration; and 2) AUD was provided with a calibration clip, that was the neutral clip of the participant. These results are summarized in Tab. II. One can observe that the second approach improved the performance, in particular, the recognition accuracy for Group 2 significantly increased from 32% to 75%.

TABLE II

EMOTION CLASSIFICATION ACCURACY (%) ON THE TEST SET.

Calibration video	Group 1	Group 2	Average
not provided	71.9	32.3	52.1
provided	73.4	75.0	74.2

We further examined which emotion was the easiest and/or the hardest to recognise. In Fig. 3, we presented the confusion matrix for Group 1 with calibration. As expected, the classifier was very successful at recognising neutral. The classifier was successful at recognising surprise (0.812) and happiness (0.750), but not disgust (0.375). This was probably due to the difficulty of expressing the disgust emotion as reported by many of our participants.

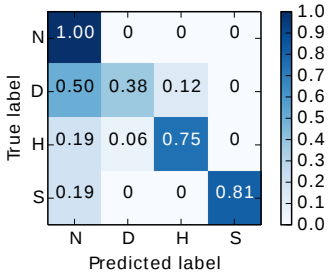

Fig. 3. Normalised confusion matrix for testing the emotion classifier on the in-house dataset (Group 1 with calibration video provided; neutral (N), disgust (D), happiness (H), surprise (S)).

C. Head Pose Filtering

We implemented the constrained-state Kalman filter in Python using the CVXOPT library [29] for the optimization (quadratic programming) problem given by (3). To estimate the measurement noise parameters R_Y and R_P for yaw and pitch angles, we used the database with labeled head pose images of Gourier et al. [30]. We measured the variance in head pose estimates from HPD for a fixed head pose (still image from database projected on another screen) sensed by the web camera. We took 1000 measurements for each of the 3 subjects at 5 different angle pairs.

For process noise estimation, we used UPNA database [31] that contains video recordings of head movements of 12 subjects (120 videos in total) labeled frame-by-frame. We fed the videos into the HPD and then into the Head Pose Filter and compared the results with the ground truth values. We searched for the process noise $\tilde{Q}$ that minimizes the sum of squared errors between outputs from the filter, $f_i(Q)$, and ground truth values, t_i, over all frames of all the videos (i.e. over all data points i). This was done separately for yaw and pitch. For this optimization problem, we used 80 % of the database. The Head Pose Filter was evaluated on the remaining 20 % of the data (6000 data points, 2 subjects) from UPNA database using the optimal parameters learned on the first 80 % of data. Fig. 4 shows the absolute errors (between filtered head pose estimates and ground truths) as they fall into different angle ranges. The absolute errors shown come from all the frames of all the videos from the held-out test set. In degrees, the mean absolute error for yaw is $4.1° \pm 3.8°$ and $2.5° \pm 2.5°$ for pitch. For real-time filtering, the whole database was used to estimate the process noises.

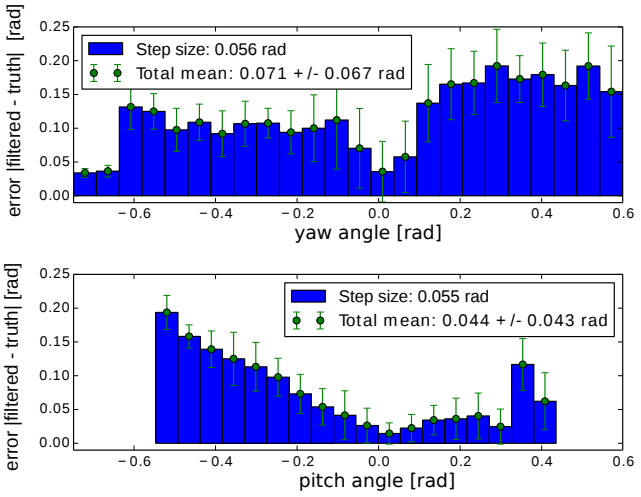

Fig. 4. Head Pose Filter evaluation on the UPNA database: absolute errors (y-axis) between filtered head pose estimates and ground truth values for different angles (x-axis), for yaw (top) and pitch (bottom).

D. User Study

The video recordings of 10 best-performing (highest testing accuracy) participants from Group 1 (with calibration video), resulting in a total of 80 clips, were replicated on the robot, as illustrated in Fig. 2. We chose the best-performing participants because the goal here was to investigate the external observers' perception of expressed emotions and head movements through the robot avatar. To achieve this, we needed to minimise the inter-subject variability of the expressions produced, and focus on whether the replication was an effective way to communicate the head and facial behaviours.

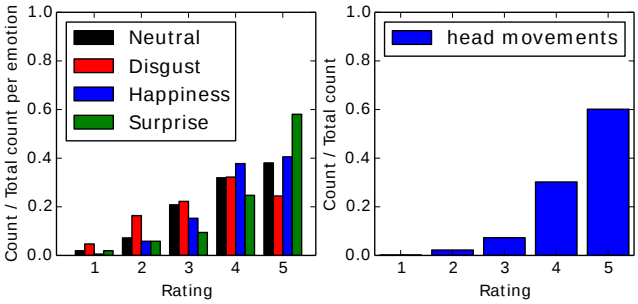

Fig. 5. Histograms of responses for replication quality of expressions of emotions (left) and head movements (right) obtained via the web survey. Ratings on the x-axis correspond to 5-point Likert scale (very poor–very good).

Using an online survey, we asked 18 external observers to watch each participant's clip along with the robot imitating their emotions and head movements, and rate the replication quality on a 5-point Likert scale. The external observers were also provided with emotion-colour mappings, as we were not concerned with how well a colour represented an emotion. Histograms of responses for emotions and head pose are shown in Fig. 5. We observed that the replication of surprise was considered most accurate, whereas disgust was not communicated that well. Neutral and happiness were judged similarly, and received better ratings than disgust.

In order to assess the observers' agreement, we calculated the intraclass correlation coefficient (ICC) [32] presented in Tab. III. According to the guidelines for interpretation of ICC measures [33], the observers' consensus would be considered *excellent* for all the target sets, except for the head pose which would be interpreted as *good*.

TABLE III
INTRACLASS CORRELATION COEFFICIENTS (ICC) FOR THE WEB-SURVEY.

Rated targets	ICC	95% CI
Head pose	0.7278	$\langle 0.6328; 0.8074 \rangle$
Neutral emotions	0.8896	$\langle 0.8047; 0.9489 \rangle$
Disgust emotions	0.9212	$\langle 0.8606; 0.9635 \rangle$
Happiness emotions	0.8021	$\langle 0.6498; 0.9083 \rangle$
Surprise emotions	0.9398	$\langle 0.8934; 0.9721 \rangle$
All emotions	0.9184	$\langle 0.8899; 0.9423 \rangle$
Head pose and all emotions	0.9055	$\langle 0.8828; 0.9256 \rangle$

E. Replication Latency

We focused on processing latency defined as the time from the frame capture to the dispatch of commands to the robot. The measured latencies were 49 ± 7 ms and 20 ± 3 ms for emotion replication and head pose replication, respectively. This was measured separately for emotions and head pose over 4.5 minutes of replication time.

V. CONCLUSION AND FUTURE WORK

In this paper, we developed a teleoperation system that replicates the facial expressions of emotions (neutral, disgust,

happiness, and surprise) and the head pose on the fly in a non-invasive manner. User evaluation obtained via a web survey shows that the proposed teleoperation system can communicate emotions and head movements very well with a high inter-rater agreement. The ratings provided for the disgust emotion suggest that this emotion is more difficult to replicate on the fly.

The results of the emotion classification suggest that the AU detector relies on a reference neutral face in order to work accurately. Classification does not rely on calibration if the recording analysed starts with a neutral face. When comparing AVG and CONCAT strategies, we observed that it was better to use the low-dimensional feature vectors generated by averaging rather than by concatenating AU information from the last W frames. From the evaluation of the head pose filter, we conclude that MAE over different angle ranges is relatively small. However, this error increases as the angle from neutral position increases.

The processing latency measurements show that the emotion replication is more computationally intensive than the head pose replication, which was expected. However, these measurements do not include the delay caused by communication over the network which would constitute the major portion of the overall delay in a teleoperation system. Thus, we can only compare the processing delay of our system with the head pose replication system proposed in [14]. They measured the overall replication latency including the network communication delay to be 100 to 170 ms and estimated the latency caused by communication over network to be 34 to 104 ms. Thus, we conclude that when compared to [14], on average we achieved a lower processing latency (by approximately 70%) for head movement replication.

As future work, we plan to extend the teleoperation system to different robotic platforms (e.g. a robot with silicon-made actuated skin).

Acknowledgements. *This work was funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref: EP/L00416X/1).*

REFERENCES

[1] O. Daly-Jones, A. Monk, and L. Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *Int. Journal of Human-Computer Studies*, 49(1):21–58, 1998.

[2] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human-Computer Interaction*, 8(4):389–428, December 1993.

[3] F. Tanaka, T. Takahashi, S. Matsuzoe, N. Tazawa, and M. Morita. Telepresence robot helps children in communicating with teachers who speak a different language. In *Proc. ACM/IEEE HRI*, 2014.

[4] Y.-S. Chen, J.-M. Lu, and Y.-L. Hsu. Design and evaluation of a telepresence robot for interpersonal communication with older adults. In *Proc. ICOST*. Springer, 2013.

[5] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.

[6] K. Kuwamura, T. Minato, S. Nishio, and H. Ishiguro. Personality distortion in communication through teleoperated robots. In *Proc IEEE RO-MAN*. IEEE, September 2012.

[7] P. Bremner, O. Çeliktutan, and H. Gunes. Personality perception of robot avatar tele-operators. In *The Eleventh ACM/IEEE HRI*, 2016.

[8] O. Celiktutan, P. Bremner, and H. Gunes. Personality classification from robot-mediated communication cues. In *Proc. IEEE RO-MAN*, 2016.

[9] P. Chevalier, J. Martin, B. Isableu, and A. Tapus. Impact of personality on the recognition of emotion expressed via human, virtual, and robotic embodiments. In *Proc. IEEE RO-MAN*, 2015.

[10] D. O. Johnson, R. H. Cuijpers, and D. van der Pol. Imitating human emotions with artificial facial expressions. *Int. Journal of Social Robotics*, 5(4):503–513, 2013.

[11] S. Song and S. Yamada. Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In *Proc. ACM/IEEE HRI*. ACM, 2017.

[12] P. Bremner and U. Leonards. Efficiency of speech and iconic gesture integration for robotic and human communicators-a direct comparison. In *Proc. IEEE ICRA*, 2015.

[13] K. Shiarlis, J. Messias, M. Someren, S. Whiteson, J. Kim, J. Vroon, G. Englebienne, K. Truong, V. Evers, N. Pérez-Higueras, et al. Teresa: A socially intelligent semi-autonomous telepresence system. 2015.

[14] P. Agarwal, S. Al Moubayed, A. Alspach, J. Kim, E. J. Carter, J. F. Lehman, and K. Yamane. Imitating human movement with teleoperated robotic head. In *Proc. IEEE RO-MAN*, 2016.

[15] P. Ekman and W. V. Friesen. Facial action coding system: A technique for the measurement of facial movement. 1978.

[16] M. Pantic. Automatic analysis of facial expressions. *Encyclopedia of Biometrics*, pages 128–134, 2015.

[17] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE CVPR*, 2013.

[18] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro. Local Zernike moment representations for facial affect recognition. In *Proc. British Machine Vision Conf.*, 2013.

[19] E. Sariyanidi, H. Gunes, and A. Cavallaro. Biologically-inspired motion encoding for robust global motion estimation. *IEEE Trans. on Image Processing*, 2017.

[20] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703, 1985.

[21] R. Shadmehr and S. P. Wise. *The computational neurobiology of reaching and pointing: a foundation for motor learning*. MIT press, 2005.

[22] D. Simon. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory & Applications*, 4(8):1303–1318, 2010.

[23] Aldebaran robotics web site. `http://www.aldebaran-robotics.com/en`, 2013.

[24] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[25] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE ICME*, 2005.

[26] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Proc. IEEE FG Workshops*, 2011.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proc. IEEE CVPRW*, 2010.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] M. S. Andersen, J. Dahl, and L. Vandenberghe. Cvxopt: A python package for convex optimization. *Available at cvxopt. org*, 2013.

[30] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, volume 6, 2004.

[31] M. Ariz, J. J. Bengoechea, A. Villanueva, and R. Cabeza. A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. *CVIU*, 148:201–210, 2016.

[32] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[33] D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994.