Face2Gesture: Translating Facial Expressions Into Robot Movements Through Shared Latent Space Neural Networks

Face2Gesture: Translating Facial Expressions Into Robot Movements Through Shared Latent Space Neural Networks*

MICHAEL SUGUITAN*, Independent Researcher, USA
NICK DEPALMA*, Plus One Robotics, USA
GUY HOFFMAN, Mechanical and Aerospace Engineering, Cornell University, USA
JESSICA HODGINS*, Robotics Institute, Carnegie Mellon University, USA

In this work, we present a method for personalizing human-robot interaction by using emotive facial expressions to generate affective robot movements. Movement is an important medium for robots to communicate affective states, but the expertise and time required to craft new robot movements promotes a reliance on fixed preprogrammed behaviors. Enabling robots to respond to multimodal user input with newly generated movements could stave off staleness of interaction and convey a deeper degree of affective understanding than current retrieval-based methods. We use autoencoder neural networks to compress robot movement data and facial expression images into a shared latent embedding space. Then, we use a reconstruction loss to generate movements from these embeddings and triplet loss to align the embeddings by emotion classes rather than data modality. To subjectively evaluate our method, we conducted a user survey and found that generated happy and sad movements could be matched to their source face images. However, angry movements were most often mismatched to sad images. This multimodal data-driven generative method can expand an interactive agent's behavior library and could be adopted for other multimodal affective applications.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies →** *Machine learning algorithms.*

Additional Key Words and Phrases: Human-robot interaction; social robots; neural networks; affective computing; behavior generation

1 INTRODUCTION

We present a method for personalizing human-robot interaction by using emotive facial expressions to generate affective robot movements. Robots can use movement to convey internal affective states for more compelling human-robot interaction. However, creating movements often requires working knowledge of robotics and kinematics. Even more accessible methods such as kinesthetic teaching are constrained by limited access to robots. Relying primarily on retrieving preprogrammed user-crafted responses from a static database can eventually diminish users' interest in the robot [25]. Generating new behaviors in response to different users' inputs may mitigate this novelty effect and promote prolonged interaction. Machine learning models, particularly deep neural networks, have achieved state-of-the-art performance in a variety of applications, such as perceived emotion recognition [38]. Neural networks have also shown promise in data generation, such as generative

*The first, second, and last authors were affiliated with Facebook AI Research when they performed this work.

Authors' addresses: Michael Suguitan*, Independent Researcher, Raleigh, North Carolina, USA; Nick DePalma*, Plus One Robotics, Pittsburgh, Pennsylvania, USA; Guy Hoffman, Mechanical and Aerospace Engineering, Cornell University, Ithaca, New York, USA; Jessica Hodgins*, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

2573-9522/2023/10-ART
https://doi.org/10.1145/3623386

adversarial networks for photorealistic images and conversational chatbots [15, 20]. Therefore, we believe that neural networks are well-suited for affective generation applications.

As a proposed application, we envision a personalized interaction scenario where a human interactant's facial expressions actively generate a robot's movement responses, e.g. in the greetings or acknowledgements. We chose to mirror facial expressions as the input modality given the industry standard of cameras installed on robots and the importance of gaze in improving subjective and social evaluations of robots [2]. We also considered the practical availability of facial expression datasets with emotion labels (e.g. the Cohn-Kanade facial expression database [23]) as well as the potential positive effects of affective mirroring in cooperative scenarios through emotional contagion [7]. We chose robot movement as the output modality given the uniqueness of this affordance to embodied robots compared to unembodied voice- or screen-based agents [18]. For a given robot, prior expert and novice users create a dataset of manually crafted movements labelled according to emotions, e.g. happiness, sadness, anger. To translate between the modalities, we propose using neural networks to learn the alignment between the facial images and the sequential movements while maintaining a semantic link through the shared emotion labels. In this scenario, the network-generated movements do not supplant, but rather *complement* the existing user-crafted movement dataset to expand the robot's available behavior library. In application, the network-generated movements act as "inbetweens" either chaining together user-crafted "key" movements or acting as idling motions.

We implemented this approach using the zoomorphic Blossom robot [40]. We used a convolutional variational autoencoder (VAE) to compress Blossom's emotion-labelled movements - head roll, pitch, yaw, and vertical translation - into a latent embedding space, and a convolutional image encoder to compress emotion-labelled facial expression images into the same latent space. To align the disparate modalities in the shared latent space, we implemented a triplet loss objective to cluster embeddings by emotion classes rather than by modality. We evaluated this approach in an online user survey where participants watched a video of a robot movement and selected the best corresponding facial expression image from a set of possible images, i.e. matching a happy movement to a happy face image. We found that generated happy and sad movements were well-matched, but angry movements were mostly mismatched to sad images.

Our contribution is an approach for translating facial expression images into affective robot movements using neural networks. Prior works in robot behavior generation synthesize new behaviors according to the robot's communicative affordances (e.g. movement, speech) [29, 30, 41]. Other works in affective human-robot interaction bisect the robot's response generation process: a classification model first recognizes human input according to discrete emotion classes, then a retrieval system selects an appropriate robot response from a predefined library of behaviors [13, 26, 31]. Compared to these works, we implement an end-to-end multimodal neural network that learns an alignment between disparate input and output modalities and can directly translate facial expressions into affectively appropriate robot movements. Our approach has further implications for expanding an agent's behavior library and for other multimodal affective applications, e.g. a listening ear responding to perceived text sentiment and audio inflection, or a video-watching companion reacting to the multimodal context of video.

2 RELATED WORK

We based our approach upon prior works in robot movement creation and neural networks.

2.1 Robot movement

Movement enables robots to interact with the world with affordances beyond screen- and audio-based agents [18]. Apart from goal-oriented actions such as locomotion or manipulation, movement can also communicate affective states, either in discrete categories (e.g. happy, sad, angry [12]) or on a continuous spectrum (e.g. valence, arousal [37]). LaViers argues that humanizing movement is of paramount importance for human-robot interaction, and

recommends referencing movement-based arts such as dance and acting in the design of robot movements [24]. However, designing emotive movements requires depth of knowledge in robotics, movement analysis, and affective expression. Learning from demonstration through either direct manipulation of a robot's actuators or remote teleoperation [3] is more accessible to lay users but still requires physical access to a robot and may not be generalizable to other platforms. To reduce the need for hand-made user-crafted behaviors, researchers have explored generating movements using machine learning models [9, 19]. We are interested in generating affective movements for robots using machine learning techniques, specifically neural networks. We view these generated movements not as supplanting the user-crafted movements, but rather complementing them to expand the robot's available behaviors.

2.2 Neural networks

2.2.1 Robot movement generation. Designing robot movements is often time-intensive and limited by proximity to physical robots. Machine learning models can use existing movements to expand a robot's available behavior library. Marmpena et al. generated motion for a humanoid robot by chaining poses together from a VAE's learned latent space [29, 30]. Yoon et al. generated gesticulation motions for a humanoid robot using a multimodal dataset of speech, text, and posture [41]. The works in this space have largely focused on humanoid embodiments, perhaps due to the familiarity and availability of humanoid movement data. Additionally, these approaches rely on datasets that are either expert-crafted or sourceable in large quantities, e.g. professionally recorded speeches to yield paired multimodal datasets. We adopt similar neural network methods, but instead rely on user-crafted movements. We believe that sourcing movements from users is a more accessible approach and yields samples that better reflect the potential end users of such a system.

2.2.2 Applications of affective movement. The ability for data-driven neural networks to learn features is useful for applications that may otherwise be intractable with heuristics, such as perceived affective recognition and generation. Many works in this space focus on perceptive tasks, such as supervised perceived sentiment analysis in text and images [11, 21]. Pakrasi et al. [34] apply notions of Kansei Engineering principles to the relationship between animated motion, the design of the character, and choice of character archetypes. Simple non-verbal movements have been shown to improve perception of team work [22] and have even been shown to improve team performance [8] in human robot teams. However connecting automatic motion generation to perceptual outcomes is still an open topic. Heimerdinger et al. [17] link context and environment to the perceived valence and arousal perceptions. However, there are still significant challenges to generating robotic motion using neural networks that are perceived affectively by people. Robots such as iCat [35], Muecas [10], and Miro [14] emphasize the affordance of facial expressions of the robot. Muecas uses computer vision to both recognize users' facial expressions and, successively, author robot facial expressions. However, this system discretizes the input user facial expression into discrete categories (neutral, happy, sad, fear, anger), from which the pre-crafted robot expressions are selected; this alignment is hand-crafted. We are interested in expanding beyond this with recent advancements in larger neural networks, which can operate end-to-end sans discretization by using the emotion labels as an alignment guide. Neural networks can also generate emotive samples of images and audio [27, 33]. Our proposed application is less technically complex than these examples, particularly in the relatively low dimensionality of the robot's movement compared to high-dimensional images, text, and audio. We show, however, that this low-dimensional space is sufficient for clustering affective states of movement.

2.2.3 Multimodal machine learning. The ability for neural networks to learn features is also useful for multimodal applications [6]. Automatic image captioning is a common application that learns alignments within a paired dataset of images and their corresponding textual descriptions [5]. Reversing the task to generate images given text descriptions is a more complex task, but recent state-of-the-art techniques are capable of generating realistic

samples [36]. Nguyen et al. adopted similar techniques to perform manifold alignment on a paired image and text dataset for robot understanding [32]. These techniques are applicable to the multimodal input-output modalities of robots, e.g. sensor inputs from cameras or microphones and movement outputs through actuators.

Prior works in affective human-robot interaction bisect robot response generation into recognition and retrieval. First, a classification model recognizes affect from human inputs (e.g. facial expression, speech) according to discrete emotion categories (e.g. happiness, sadness, anger) [13, 26, 31]. The system then uses the recognized emotion to retrieve an appropriate robot response from a predefined library of behaviors.

In our approach, we bypass the intermediate classification step by using an end-to-end multimodal neural network with an encoder-decoder architecture. The network aligns the disparate input and output modalities by using the emotion labels to structure its latent embedding space. In the embedding space, the network clusters data together towards similar labels and away from opposing labels. To generate a behavior, the network encodes the input into the embedding space and decodes into an output with the same emotion class. Our approach directly translates the inputs into affectively appropriate outputs and generates behaviors beyond the initial library.

We explored neural network-based techniques for robot movement generation in prior work titled "MoveAE" [39]. In that work, we used an earlier subsection of Blossom's movement dataset in a movement-only VAE for movement generation and affective modification. After training the VAE to compress the movement data into a latent space, we generated and modified movements. To generate movements, we sampled embeddings in the latent space and decoded through the latter decoder half of the VAE into new network-generated movements. To modify a movement, we first selected a base movement with a given emotion label and encoded through the former encoder half of the VAE into a latent embedding. Beside the network, we used linear regression to map the latent space into the 2D circumplex model of affect, a relational representation of emotions on a 2D plane of valence and arousal dimensions [37]. In this valence-arousal plane, we moved embeddings from their original labeled emotion to a new target emotion, e.g. moving an originally happy-labeled movement embedding to the sad region in the valence-arousal space. We then decoded the embedding through the latter decoder half of the VAE into a network-generated affect-modified movement.

The network presented here, dubbed "Face2Gesture," expands upon the movement-only MoveAE by translating affective facial expressions into robot behaviors. MoveAE implemented modification within the modality of movement; Face2Gesture implements translation between the input modality of images and output modality of movement. Face2Gesture builds upon MoveAE by:

- Refactoring and restructuring the VAE neural network,
- Using a larger movement dataset,
- Implementing a paired image-based VAE network for the facial expressions from the Cohn-Kanade database [23],
- Optimizing on a triplet loss to align the disparate movement and image modalities in a shared latent space, and
- Outputting either reconstructions of user-crafted movements or newly synthesized image-generated movements.

We use techniques from these previous works to create an affective response system that generates robot movements from facial expressions. We perform intermodal translation by using techniques from multimodal machine learning, specifically encoder-decoder architectures and emotion label-based triplet loss. The resulting system encodes both robot movements and human facial expressions into a shared latent embedding space, and decodes these embeddings to generate movements from either modality.

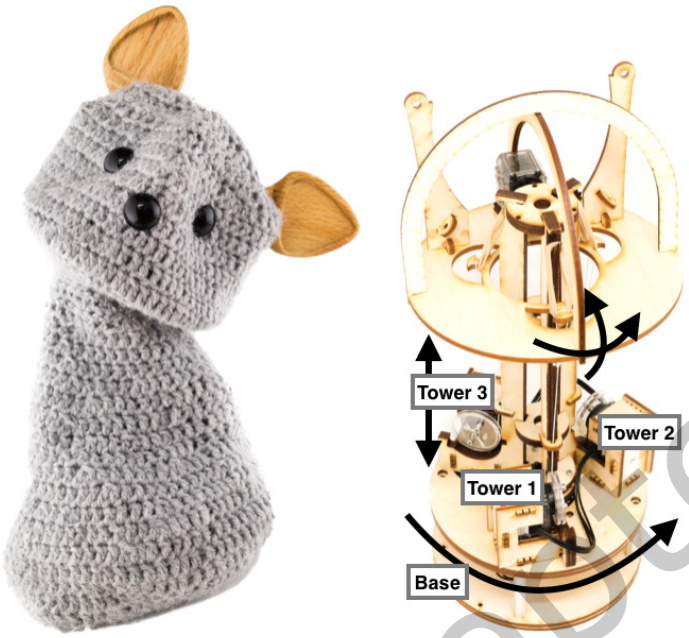

Fig. 1. The Blossom robot. The exterior (left) is made of soft materials while the interior mechanism (right) consists of a central tower structure from which the head platform is suspended by elastic bands. The head platform has four degrees of freedom: roll, pitch, yaw, and vertical translation.

3 METHODS

We used an existing robot platform, datasets of movements and face images, and encoder-decoder neural networks.

3.1 Robot platform

We used the Blossom robot, an open-source social robot (Figure 1) [40]. Blossom's internal mechanisms consist of a head platform suspended from a tower structure that rotates about its base platform. Blossom features four degrees of freedom (DoFs): roll, pitch, yaw, and vertical translation, though we disable vertical translation to simplify the control interface. The robot achieves motion with four actuators: tower motors 1, 2, and 3 control the front, left, and right sides of the head, respectively, and a motor in the base rotates the tower left and right. The robot's head can pitch up and down, roll left and right ±45°, yaw left and right ±150°about its base, and vertically translate up and down. Although the robot's DoFs are limited compared to more complex embodiments, it features a large range of motion and head movements alone can convey complex affective information [1]. Users can control the robot with a mobile browser-based application that maps the orientation of the phone into motion for the robot's body.

3.2 Data

3.2.1 Movements. We used robot movement samples that we crowdsourced from lay users. We asked users to first view video prompts of cartoon characters (SpongeBob, Pikachu, Homer Simpson - recognizable characters with recognizable facial and bodily emotional expressions) conveying different emotions (happiness, sadness, anger), then to puppeteer the robot with their phones as if it were conveying the same emotion. Some movements

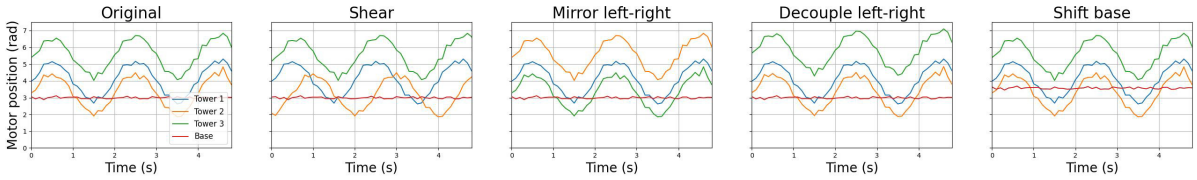

Fig. 2. Examples of Blossom movement data and dataset augmentations visualized through motor trajectories. Tower 1 controls the pitch of the front of the head, towers 2 and 3 control the left-right rolling of the head, and base controls left-right yaw. Each original movement is a 4.8-second sample from a user-crafted movement (left). The horizontal axis is time; the vertical axis is the radial motor position. Shearing the degrees of freedom (DoFs) in time slightly nudges their trajectories relative to each other. Mirroring horizontally swaps the left and right tower motors (2 and 3) and reverses the base rotation. Decoupling the left and right tower motors separates the DoFs to promote rolling motion. Shifting the average base rotation slightly promotes yawing motion.

were collected locally in-person, though most were collected remotely by users teleoperating the robot. Users generally found anger the hardest emotion to convey. To account for the subjectivity of the user-crafted samples, we filtered the dataset by deploying a survey to another set of users. Each question contained a video of the robot performing each movement, followed by a question asking users to select the conveyed emotion. We deployed this filtering survey through Amazon Mechanical Turk and received over 250 responses, averaging 25 ratings for each movement. We kept only movements recognized at a threshold of 50%, an arbitrary margin above the chance level of 33% for each of the three emotions. This filtering downsized the original dataset from over 200 movements samples to approximately 140 movement samples. We then balanced the emotion classes by oversampling from the smaller class populations. Because the neural network requires fixed-length inputs, we took random 4.8-second samples from each movement. Though we can expand the data through augmentation, we took care to perform only augmentations that are emotionally neutral, e.g. mirroring a movement from left to right is neutral and valid, but modulating the pitch of the robot's head downwards or upwards may affect its conveyance of sadness and is thus invalid. We designed the following augmentations (Figure 2):

- **Shearing** the DoFs in time by slightly nudging their trajectories relative to each other.
- **Mirroring** horizontally (i.e. along a vertical plane bisecting the left and right halves of the robot) by swapping the left and right tower motors (2 and 3) and reversing the base rotation.
- **Decoupling** the left and right tower motors. Because these motors are often synchronized in the user-crafted movements, they have a tendency to collapse into copies of each other. Separating these DoFs slightly promotes rolling motion without modifying the emotion.
- **Shifting** the average base rotation slightly. Because the robot faces directly forward in many user-crafted movements, this augmentation compensates for the neglect of the base motor and promotes yawing motion.

Because of the relatively small size of the user-crafted movement dataset, enabling augmentation was necessary to avoid completely overfitting.

3.2.2 Face images. We used the Cohn-Kanade dataset, a collection of facial expression videos from a diverse range of actors [23]. We used the final frame at the apex of each emotion, resulting in approximately 150 samples. We augmented the data with low-magnitude rotation, translation, horizontal mirroring, scale, shear, and brightness transformations.

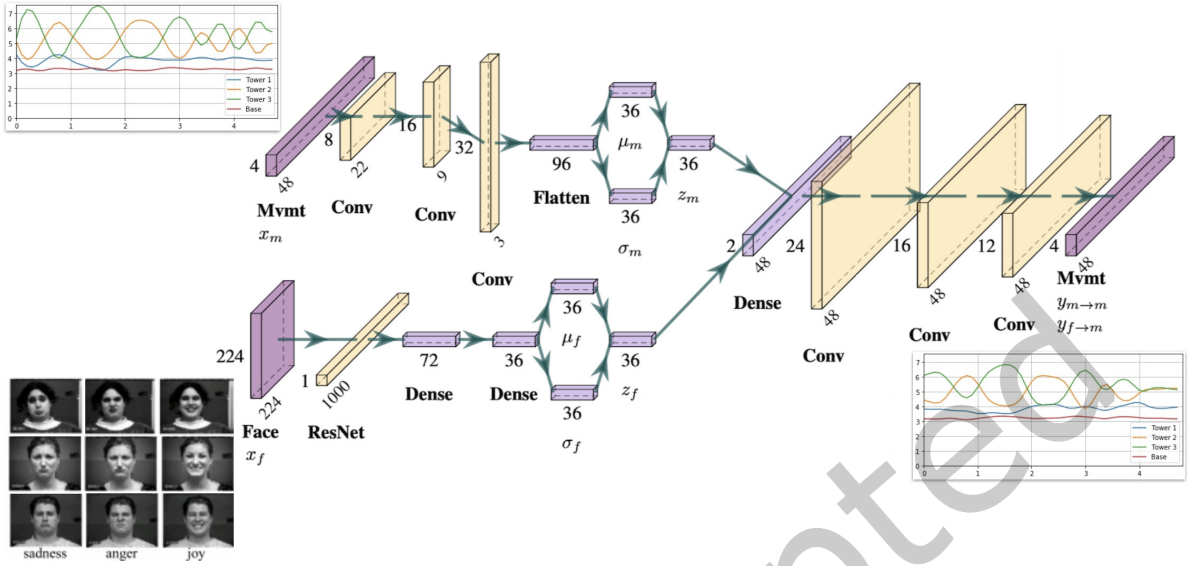

Fig. 3. Neural network for translating face images into movements. The user-crafted ***m***ovements x_m (4.8 seconds at 10 Hz with four DoFs $\rightarrow 48 \times 4$) are encoded into a 36D embedding space $z_m \sim N(\mu_m, \sigma_m)$ (top left). The movement embeddings z_m are then decoded to reconstruct the original input $y_{m\to m}$ (right). The ***f***ace images x_f are encoded into the same 36D embedding space $z_f \sim N(\mu_f, \sigma_f)$ (bottom left). The face embeddings z_f are then decoded to generate new movements $y_{f\to m}$ (right). *(Face images ©Jeffrey Cohn).*

Algorithm 1: Training algorithm

Input : Input movements X_m, input face images X_f

$F_m(x_m) \leftarrow f_{dec}(f_{enc}(x_m))$ `//movement autoencoder neural network;`

$F_f(x_f) \leftarrow f_{embd}(ResNet_{50}(x_f))$ `//face image encoder neural network;`

while *not converged* **do**

  x_m, x_f `//minibatch of movements and faces;`

  $y_{m\to m} \leftarrow F_m(x_m)$ `//movement reconstructions;`

  $z_m \leftarrow f_{enc}(x_m)$ `//movement embeddings;`

  $z_f \leftarrow F_f(x_f)$ `//face embeddings;`

  $L_r \leftarrow MSE(y_{m\to m}, x_m)$ `//reconstruction loss with mean-squared error;`

  $L_{KL,m} \leftarrow KL(z_m)$ `//movement KL divergence;`

  $L_{KL,f} \leftarrow KL(z_f)$ `//face KL divergence;`

  $L_t \leftarrow T(z_m, z_f)$ `//triplet loss (Equation 1);`

  $L \leftarrow w_r L_r + w_{KL,m} L_{KL,m} + w_{KL,f} L_{KL,f} + w_t L_t$ `//overall loss, backpropagate to update networks` F_m `and` F_f`;`

  $y_{f\to m} \leftarrow f_{dec}(z_f)$ `//pass face embeddings through decoder to generate movements;`

end

$F_{f\to m}(x_f) \leftarrow f_{dec}(F_f(x_f))$ `//face-to-movement translation network;`

3.3 Network

We constructed the end-to-end network using convolutional encoders and decoders for each data modality. We aligned the encoded latent spaces using triplet loss.

3.3.1 Movement VAE. We used a VAE to compress the movement data into embeddings in a lower-dimension latent space (Figure 3, top left to right). The encoder f_{enc} uses 1D convolutions that stride across the time dimension of the ***m***ovements $x_m \in X_m$, and outputs the latent space distribution parameters (log-mean and log-variance of a distribution $N(\mu_m, \sigma_m)$). We empirically set the latent dimension to 36 parameters; we arrived at this dimensionality by decreasing the latent space size until the reconstructed movements lost too much information, i.e. were very smoothed out. The decoder f_{dec} uses these parameters to sample embeddings $z_m \sim N(\mu_m, \sigma_m)$ which pass through deconvolutional layers to reconstruct the original movements $y_{m \to m}$. We used LeakyReLU ($\alpha = 0.1$) and batch normalization after each convolutional and fully connected layer. We calculated the reconstruction loss L_r as the mean-squared error between the raw trajectories of the original and reconstructed movements. The VAE also uses Kullback-Leibler (KL) divergence as a loss $L_{KL,m}$ to ensure that the embedding distribution approximates a normal distribution, i.e. $N(\mu_m, \sigma_m) \approx N(0, 1)$.

3.3.2 Face image encoder. We encoded the images of ***f***aces $x_f \in X_f$ into the same latent space by first passing them through a pretrained ResNet_{50} model [16], then through two fully connected layers (Figure 3, bottom left). Similar to the VAE, we used LeakyReLU and batch normalization after the fully connected layers, and the final encoder layers yield the embedding distribution $z_f \sim N(\mu_f, \sigma_f)$. We added the KL divergence of the face embeddings $L_{KL,f}$ to the overall loss.

3.3.3 Shared latent space alignment using triplet loss. Because we do not have paired alignment between robot movements and face images, we used triplet loss L_t to align the embeddings Z_m and Z_f in the shared latent space [32]. The triplet loss minimizes the distance between an ***a***nchor embedding z_a and a positive sample embedding z_+, and maximizes the distance between the anchor and a negative sample embedding z_-. For each sample in a minibatch, we mined positive samples by randomly sampling embeddings that share the same emotion class, and negative samples from the other classes. We used an imbalanced mining scheme wherein movement embedding anchors can sample from either modality, while face embedding anchors only select positive samples from the movement embeddings. The intuition is that the image encoder can easily separate the emotions due to the pretrained ResNet_{50} model and should primarily be fine tuned to match the movement embedding space. For example, given a happy movement as an anchor, positive samples come from happy movements and images, and negative samples come from the set of sad and angry movements and images. However, given a happy face image as an anchor, positive samples come only from happy movements. We used the Euclidean distance function $d(a, b)^2$ with no margin.

$$L_t = \sum_{z_a \in Z_m \cup Z_f} max(d(z_a, z_+)^2 - d(z_a, z_-)^2, 0) \tag{1}$$

The overall loss objective of the network is a weighted combination of the reconstruction, KL, and triplet losses:

$$L = w_r L_r + w_{KL,m} L_{KL,m} + w_{KL,f} L_{KL,f} + w_t L_t \tag{2}$$

We empirically set the weights as $w_r = 1 \times 10^4$, $w_{KL,m} = 1 \times 10^{-2}$, $w_{KL,f} = 1 \times 10^{-1}$, and $w_t = 1 \times 10^3$. We adjusted these weights based on subjectively appraising the reconstructions and visually checking the clusters in the latent space.

Algorithm 1 describes the training loop. Due to the subjectivity of the outputs, we both monitored the loss curves and appraised the quality of the image-generated movements during training. After training, we can use

the function $F_{f \to m}(x_f) = f_{dec}(F_f(x_f))$ - the pipeline of the face encoder and the movement decoder - to translate face images into movements $y_{f \to m}$ (Figure 3, bottom left to right). We trained for 1,500 epochs with a learning rate of 1×10^{-2}, batch size of 32, Adam optimizer, and an 80-20 train-test split.

4 EVALUATION

We evaluated the approach through both objective technical metrics and a subjective user survey.

4.1 Network evaluation

We evaluated the technical performance of the method through its performance in minimizing the loss objectives. We also monitored the outputs: the reconstructed and image-generated movements, and the separability of the latent embedding space. As an ablation study, we analyzed the performance of the network optimizing either only reconstruction loss or only triplet loss.[1]

4.2 User evaluation

Due to the subjective nature of the proposed method's outputs, we performed a user evaluation through an online survey. We constructed a survey where each question showed a video of a movement and a lineup of three facial expression images, consisting of the movement's actual source image and two random images sampled from the other emotion classes. We asked users to view the video and select the image that best corresponds to the movement. We defined a baseline as using a source face image's known emotion label and randomly selecting a user-crafted movement sample of the same corresponding emotion class, e.g. pair a randomly chosen happy face image with a randomly chosen happy movement sample. Rather than claim that our method improves upon the baseline, our method avoids the repetitiveness of recycling a predefined library of behaviors, the benefits of which would require a longitudinal evaluation. Our simpler hypothesis is that the image-generated movements will be recognized above the 50% level used to filter the dataset (Section 3.2.1). We deployed this comparison survey on Amazon Mechanical Turk and received responses from 50 participants, each of whom viewed the same set of 30 selected user-crafted and generated movement samples.

5 RESULTS

We analyzed the results through objective technical metrics and the subjective user evaluation.

5.1 Network training

We monitored the reconstruction and triplet losses during training (Figure 4). There is a gap between the triplet training and testing loss, indicating overfitting. As explained later, this gap may be a limitation of the network's ability to separate happy and angry movements, particularly those it may not have trained on.

5.1.1 Reconstruction. We evaluated reconstruction quality by comparing the inputs x_m to the outputs $y_{m \to m}$ (Figure 5). The outputs capture the overall trajectories of the inputs, but have difficulty preserving exaggeration and tend to smooth out low-amplitude high-frequency "jittering."

5.1.2 Embedding separation. We evaluated embedding separability by visualizing the latent space $Z_m \cup Z_f$ using t-SNE (Figure 6, left) [28]. Happy and sad samples are well-aligned, but angry movements are barely separated from happy movements. This coupling may be due to the ambiguity in the data itself (i.e. happy and angry are both high arousal affective states, and are thus difficult to delineate with a simple embodiment), and may also

[1]Because KL divergence only helps shape the learned latent space but does not by itself generate movements or align embeddings, we do not ablate for a KL-only configuration.

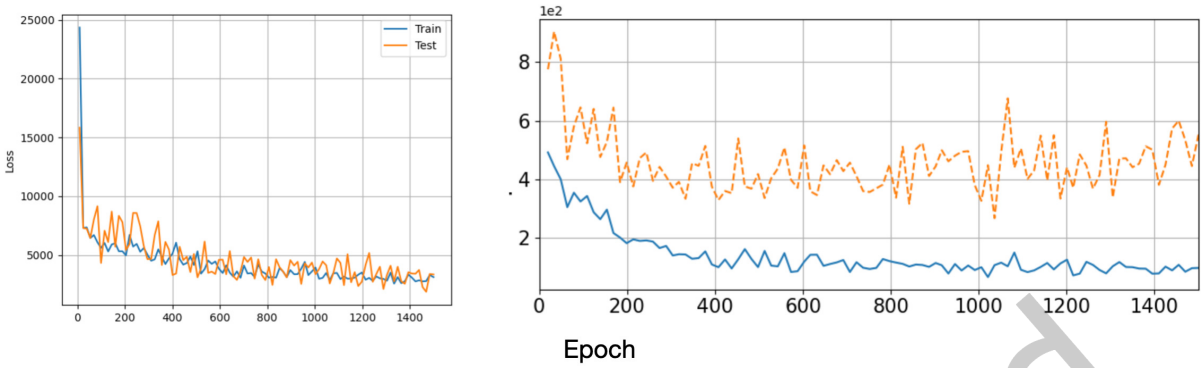

Fig. 4. Network training curves for reconstruction (left) and triplet loss (right). Triplet loss shows signs of overfitting, perhaps due to a coupling of perceptually similar happy and angry movements.

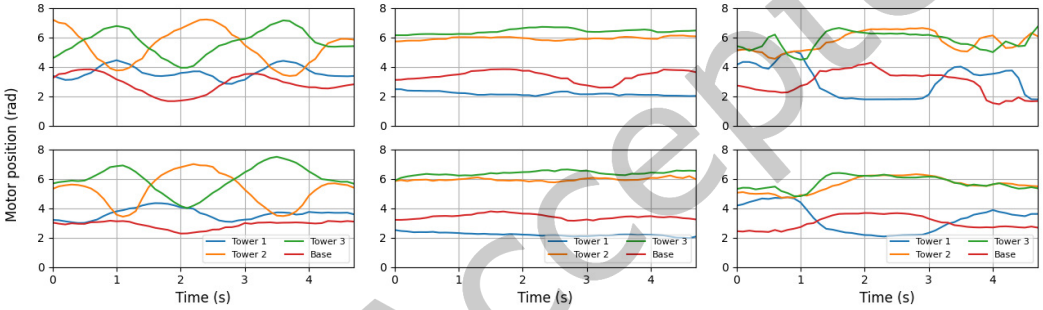

Fig. 5. Examples of original movements x_m (top) with their reconstructions $y_{m \to m}$ (bottom) (happy left, sad middle, angry right). The reconstructions maintain the overall trajectories but have difficulty preserving the exaggeration and low-frequency high-amplitude components of the originals.

explain the overfitting in the triplet loss training curve (Figure 4, bottom). Additionally, even user-crafted happy and angry movements were less correctly recognized than sad in the video user survey (Figure 11).

5.1.3 Ablation. Using only reconstruction loss defines an upper bound for generating realistic movements, but does not yield noticeable improvements (Figure 7). Addressing the deficiencies of the reconstructions (oversmoothed, limited exaggeration) may require alternate techniques such as frequency-domain representation [4, 42].

Using only triplet loss defines an upper bound for the latent space separability (Figure 6, right). Even without other objectives, angry and happy movements are still close, suggesting that the coupling is not due to the other losses, but is rather a limitation of the model itself.

5.1.4 Generation. Throughout training, we appraised the subjective quality of image-generated movements $y_{f \to m}$ (Figures 8, 9). The generated movements retain many of the characteristics of the user-crafted movements, e.g. happy movements have high tower 1 position and sinusoidal out-of-phase rolling motion in tower motors 2 and 3, sad movements have lower tower 1 position and overall flatter motion. As with the reconstructions, the generated movements have less exaggeration and jittering than the originals.

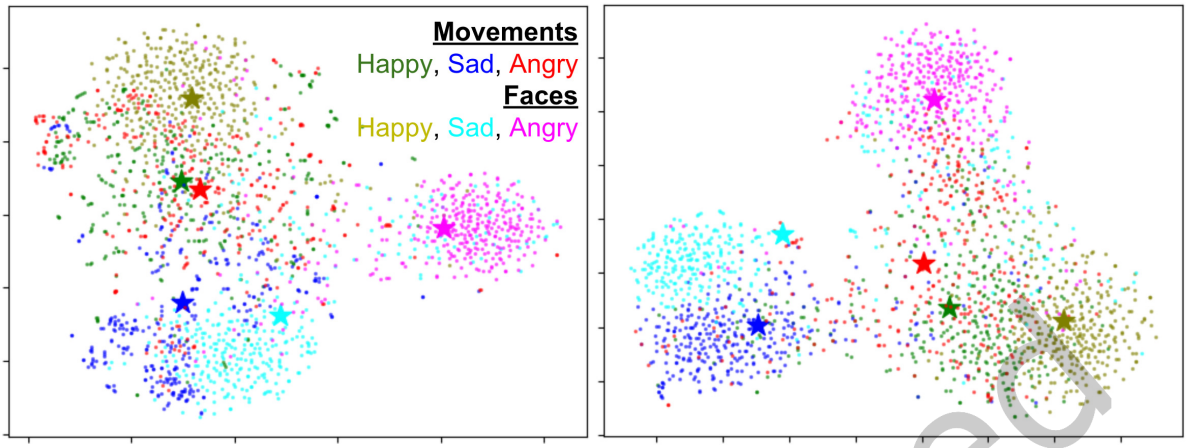

Fig. 6. t-SNE plots of the shared latent embedding space for the full multi-objective network (left) and a network optimizing only triplet loss (right). Colors indicate modality (movements, faces) and emotion (happy, sad, angry). Stars indicate centroids of each class. Happy and sad movements and faces are closely aligned, but angry movements are barely separated from happy movements, even when optimizing only for triplet loss (right).

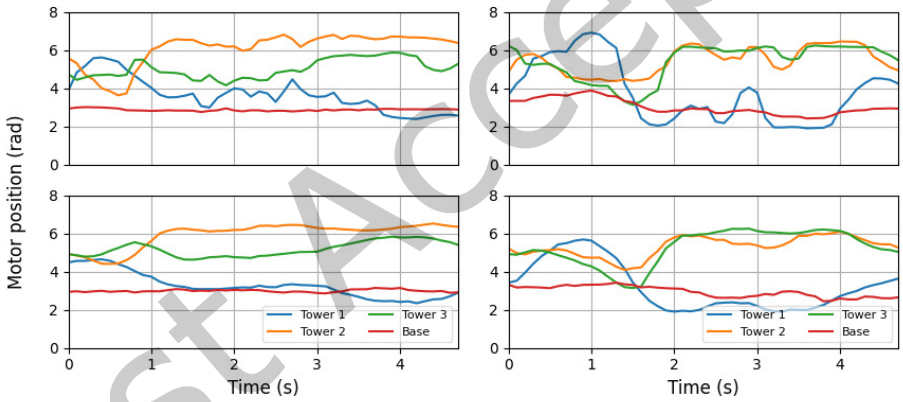

Fig. 7. Reconstructions from a network optimizing only reconstruction loss. There is only marginal improvement over the standard network (Figure 5); exaggeration is better preserved, but jittering is still smoothed out.

5.1.5 Kinematic comparison. We compared the user-crafted and image-generated movements from their respective test sets by calculating kinematic features (Table 1, Figure 10). We calculated range and speed as the peak-to-peak distance and gradient for each DoF, respectively. We calculated pitch as the difference between the positions of the front of the head (tower motor 1) and the average of the sides of the head (tower motors 2 and 3). Positive pitch is looking upwards, and negative pitch is looking downwards. We averaged speed and pitch across the length of each movement. The image-generated movements are mostly comparable to the user-crafted movements, though the user-crafted movements have larger between-class variation (Table 1, μ columns), such as the range and speed of the tower motors (Figure 10, left column). User-crafted angry movements in particular exhibit noticeably higher base range and speed than their image-generated counterparts (Figure 10, right column).

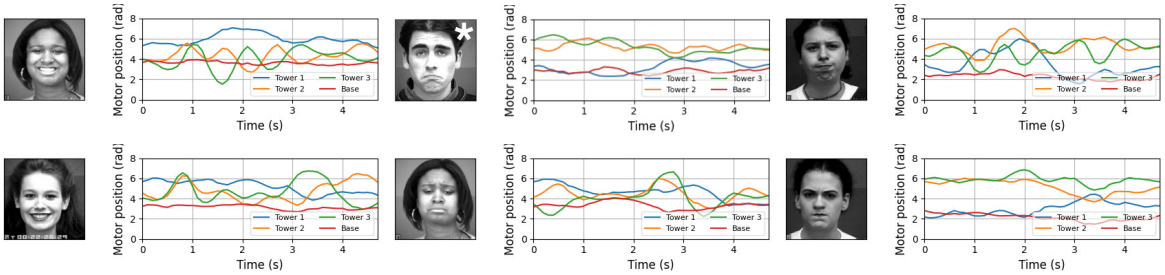

Fig. 8. Examples of source face images x_f paired with their generated movements $y_{f \to m}$ (happy left, sad middle, angry right). The generated movements maintain similar characteristics of the original user-crafted movements x_m (Figure 5), e.g. happy movements have high tower 1 position and sinusoidal out-of-phase rolling motion in tower motors 2 and 3, sad movements have lower tower 1 position and overall flatter motion. *Note: due to licensing restrictions, the face image in the second set of results (top row center) has been replaced with a publishable image from the same emotion class. (Face images ©Jeffrey Cohn).*

Table 1. Analytical comparison of the kinematic features (Figure 10). The image-generated movements approximate the trends of the mean speed, μ, of the user-crafted movements, but often have smaller standard deviations σ.

Feature	Source	Happy		Sad		Angry	
		μ_H	σ_H	μ_S	σ_S	μ_A	σ_A
Tower range	User	0.61	0.18	0.99	0.26	0.89	0.30
	Gen	0.79	0.18	0.85	0.18	0.78	0.17
Base range	User	0.31	0.23	0.26	0.14	0.92	0.65
	Gen	0.47	0.23	0.30	0.09	0.35	0.16
Tower speed	User	1.60	0.61	0.80	0.25	1.94	1.25
	Gen	1.61	0.25	1.42	0.33	1.56	0.27
Base speed	User	0.54	0.51	0.32	0.21	1.19	0.40
	Gen	0.58	0.13	0.58	0.09	0.62	0.19
Posture	User	-0.11	0.83	-2.19	0.64	-0.53	1.47
	Gen	0.92	1.04	-1.38	0.90	-1.09	0.97

5.2 User evaluation

The user evaluation serves as a subjective appraisal of the generated movements. We distributed a survey asking users to match a video of a movement - either user-crafted or image-generated - to its corresponding source facial expression image; we received 50 responses, but did not record demographic information. For the survey, we used only data from the respective movement and image test sets, i.e. samples that the network did not train on. For the user-crafted movements, we randomly paired face images only with movements from the movement test set. For the image-generated movements, we used only movements generated from images from the image test set. We used five movements for each condition, resulting in a total set of 30 movements (2 sources × 3 emotions × 5 samples). We analyzed the user evaluation results with a confusion matrix (Figure 11); perfect

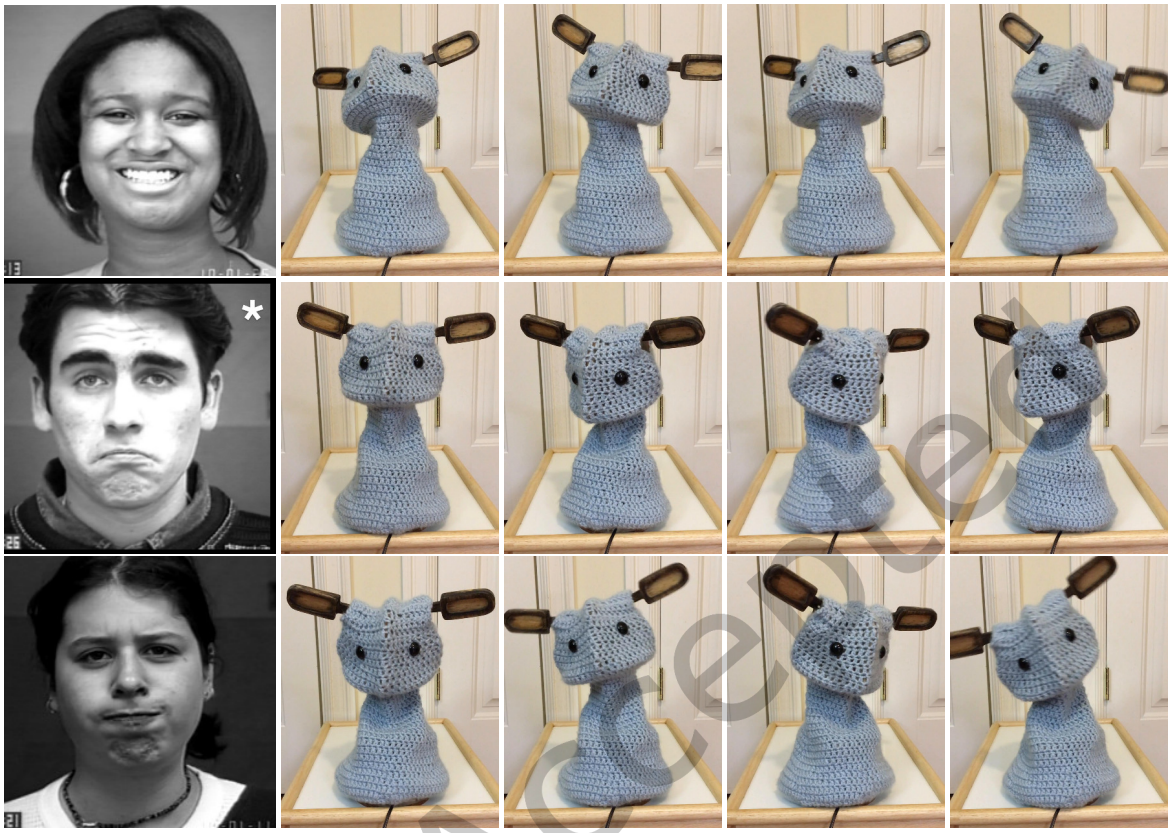

Fig. 9. Examples of image-generated happy (top), sad (middle), and angry (bottom) movements shown in the survey. * *Note: as with the previous figure, due to licensing restrictions, the face image in the second set of results (middle row) has been replaced with a publishable image from the same emotion class. (Face images ©Jeffrey Cohn).*

results would be an identity matrix. The randomly sampled user-crafted movements are overall well-matched (left). The image-generated happy and sad movements are less well-matched (right), but are still above the 50% level we used for filtering the dataset (Section 3.2.1). However, generated angry movements are recognized below chance, being confused primarily for sadness, but also for happiness. To compare the perceived recognition accuracies between the user-crafted and image-generated movements, we performed equivalence tests (two one-sided t-tests) with an equivalence bound of ±10%. These tests yielded p-values of 0.39, 0.96, and 0.99 for happy, sad, and angry, respectively, showing that none of the classes are significantly equivalent.

6 DISCUSSION

The network training results show that the network is capable of reconstructing the original user-crafted movements and generating new movements from the shared latent space. The difficulty in separating angry movements can be attributed to the limitations of both the model and the platform. Users who created movements noted that it was difficult to convey anger in particular due to the robot's lack of appendages. This limitation may have resulted in angry and happy movements being perceptually similar, as they are both classified as high-arousal emotions on the circumplex model [37]. Additionally, due to the human uninterpretability of the

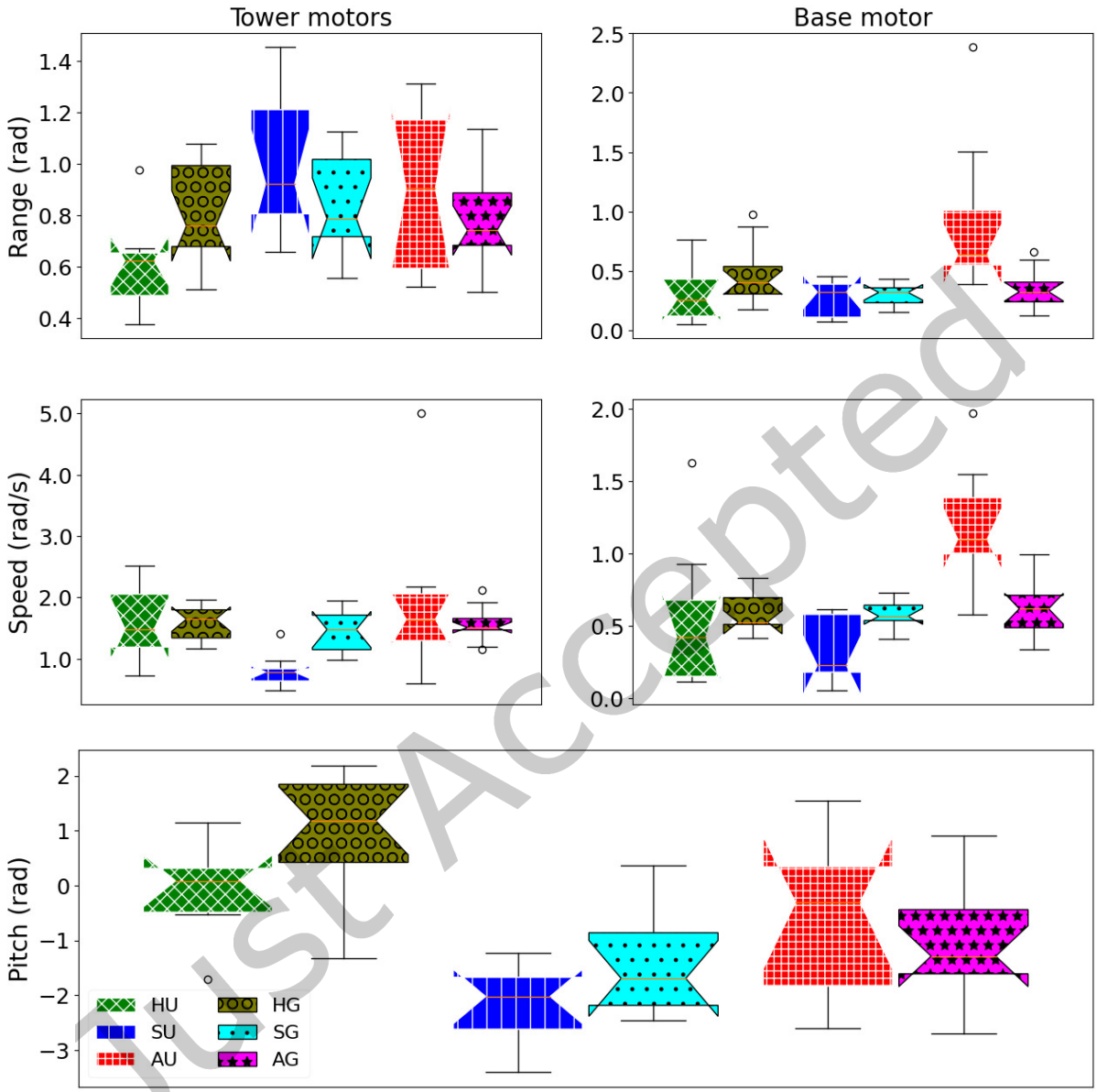

Fig. 10. Comparison of kinematic features between the user-crafted and image-generated movements. The legend (bottom left) is the emotion (**H**appy, **S**ad, **A**ngry) and source (**U**ser-Crafted, Image-**G**enerated). The user-crafted movements show more between-class variation, but the generated movements preserve many of the overall features.

learned embedding feature space and stochastic nature of t-SNE, the 2D visualization may have found more variance in latent features related to arousal and not valence, which could have delineated happy and angry samples.

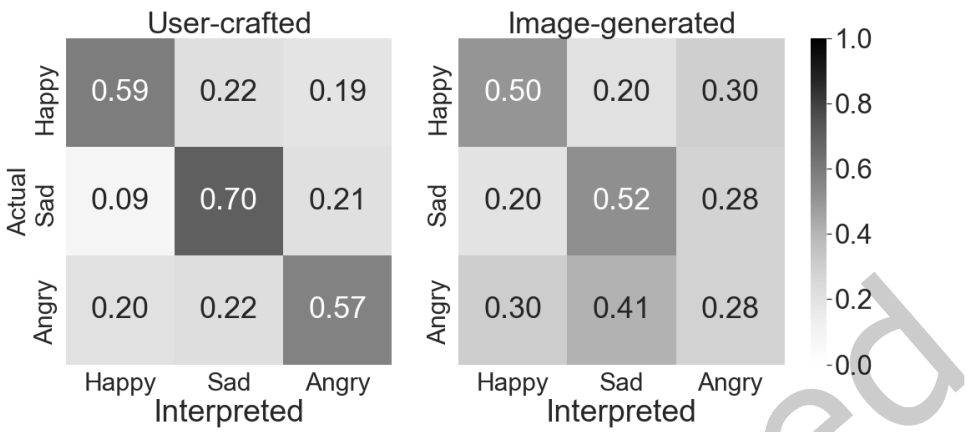

Fig. 11. Confusion matrices for both the user-crafted (left) and image-generated movements (right). Participants viewed videos of the movement then selected the best corresponding face image from a lineup. While the perceived recognition accuracies for the image-generated movements are lower, happy and sad are still recognized above the 50% level. However, generated angry movements are recognized below chance and are most often mismatched to sad images.

The confusion of generated angry movements as sad may be attributed to the difficulty in maintaining the exaggeration of the user-crafted movements, as corroborated by the kinematic analysis (Figure 10, right column). This suggests that exaggeration is an important feature for conveying anger. Though the generated happy and sad movements were recognized above chance, the accuracies were not significantly equivalent to the user-crafted movements. We view the generated movements as not supplanting, but rather complementing existing user-crafted behavior libraries. For example, an agent could use the more legible user-crafted behaviors for "active" scenarios such as call-and-response, while using the generated behaviors for "passive" scenarios such as greeting or "inbetween" motions chaining together "key" sequences. To avoid using potentially confusing generated movements (e.g. generated happy movements potentially interpreted as sad), we could filter usable generated movements by measuring similarity to the user-crafted movements (e.g. minimizing embedding distances in the latent space) or develop improved network architectures that better preserve the original movement affects.

6.1 Limitations and future work

We used only a subset of the six canonical emotions [12], which themselves are a discretization of the broad continuous spectrum of emotions [37]. This simplification was done in part to reduce the task to the most legible emotions, but also due to the limitations of the limbless and potentially velocity-constrained robot. Additionally, there may be ambiguity within the image dataset itself. Angry and sad images are both low-valence emotions that may be confounding depending upon both the performer and interpreter of the expression. This discrepancy is orthogonal to the confusion between angry and happy movements, and highlights disparities between movement and images as affective modalities. Future work could involve using a more expressive platform with more DoFs, expanding the range of emotions and data modalities (e.g. text, audio), and deploying the system in a real-time interactive scenario.

While we achieved good survey results using a between-class lineup, i.e. one image for each of the three emotion classes, the unpaired nature of the different dataset modalities would make it difficult to discern the

source image from a within-class lineup, e.g. it would be difficult to confidently select the source happy image from a lineup consisting of only happy images. Although the usability of this approach on unpaired and separately collected data can be seen as a feature, future work would benefit from collecting a paired dataset of prompts and multimodal behavior demonstrations in an attempt to achieve a deterministic translation function. Additionally, new transformer-based neural networks have achieved new state-of-the-art performance on multimodal tasks such as text-to-image generation [36]; such architectures may prove to be invaluable for future applications in affective computing, but are also increasingly complex compared to the VAE network we presented here.

7 CONCLUSION

In this work, we demonstrated an approach for generating robot behaviors from emotive images using neural networks. We used convolutional encoders to compress affective robot movements and facial expression images into a shared latent embedding space. We used a triplet loss objective to align the multimodal embeddings by emotion, e.g. bringing happy movements closer to other happy movements and faces, and separating them from sad and angry movements and faces. We then used a convolutional decoder to generate movements from embeddings from either modality. Through a subjective user evaluation, we found that happy and sad image-generated movements were recognizable and well matched to their source images above a 50% level, but generated angry movements were mostly mismatched to sad images. Though the perceived recognition accuracies were not significantly equivalent to the user-crafted movements, the generated movements are still usable for expanding the agent's behavior library. Future behavior systems for affective agents can adopt this intermodal approach with different modalities, such as generating movements from speech or other emotion-labeled inputs.

8 ACKNOWLEDGEMENTS

This work was performed while the first, second, and last authors were employed by Facebook AI Research.

REFERENCES

[1] Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2015. Decoupling facial expressions and head motions in complex emotions. In *2015 International Conference on Affective Computing and Intelligent Interaction*. 274–280. https://doi.org/10.1109/ACII.2015.7344583

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63. https://doi.org/10.5898/JHRI.6.1.Admoni

[3] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (2009), 469–483. https://doi.org/10.1016/j.robot.2008.10.024

[4] Mattia Atzeni and Diego Reforgiato Recupero. 2018. Deep learning and sentiment analysis for human-robot interaction. In *European Semantic Web Conference*. Springer, 14–18.

[5] Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing* 311 (2018), 291–304. https://doi.org/10.1016/j.neucom.2018.05.080

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[7] Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly* 47, 4 (2002), 644–675.

[8] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713.

[9] Sarah Jane Burton, Ali-Akbar Samadani, Rob Gorbet, and Dana Kulić. 2016. Laban movement analysis and affective movement generation for robots and other near-living creatures. In *Dance Notations and Robot Motion*. Springer, 25–48.

[10] Felipe Cid, Jose Moreno, Pablo Bustos, and Pedro Núñez. 2014. Muecas: A multi-sensor robotic head for affective human robot interaction and imitation. *Sensors* 14, 5 (2014), 7711–7737. https://www.proquest.com/scholarly-journals/muecas-multi-sensor-robotic-head-affective-human/docview/1537486387/se-2

[11] Cícero dos Santos and Maíra Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 69–78. https://aclanthology.org/C14-1008

[12] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (1992), 169–200. https://doi.org/10.1080/02699939208411068

[13] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. 2019. Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4011–4018. https://doi.org/10.1109/LRA.2019.2930434

[14] Moojan Ghafurian, Gabriella Lakatos, and Kerstin Dautenhahn. 2022. The zoomorphic Miro robot's affective expression design and perceived appearance. *International Journal of Social Robotics* 14 (2022), 945–962.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Madison Heimerdinger and Amy LaViers. 2019. Modeling the interactions of context and style on affect in motion perception: stylized gaits across multiple environmental contexts. *International Journal of Social Robotics* 11 (2019), 495–513.

[18] Guy Hoffman and Wendy Ju. 2014. Designing robots with movement in mind. *Journal of Human-Robot Interaction* 3, 1 (Feb. 2014), 91–122. https://doi.org/10.5898/JHRI.3.1.Hoffman

[19] Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics* 35, 4 (2016), 138.

[20] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Web, Artificial Intelligence and Network Applications*. Springer International Publishing, Cham, 946–956.

[21] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. 2019. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters* 120 (2019), 69–74. https://doi.org/10.1016/j.patrec.2019.01.008

[22] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: Easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. 1555–1566.

[23] Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*. 46–53. https://doi.org/10.1109/AFGR.2000.840611

[24] Amy LaViers. 2019. Make robot motions natural. *Nature* 565, 7740 (2019), 422–424.

[25] Iolanda Leite, Carlos Martinho, Andre Pereira, and Ana Paiva. 2009. As time goes by: Long-term evaluation of social presence in robotic companions. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. 669–674. https://doi.org/10.1109/ROMAN.2009.5326256

[26] Tzuu-Hseng S. Li, Ping-Huan Kuo, Ting-Nan Tsai, and Po-Chien Luan. 2019. CNN- and LSTM-based facial expression analysis model for a humanoid robot. *IEEE Access* 7 (2019), 93998–94011. https://doi.org/10.1109/ACCESS.2019.2928364

[27] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018). https://ojs.aaai.org/index.php/AAAI/article/view/11955

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using T-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[29] Mina Marmpena. 2021. *Emotional body language synthesis for humanoid robots*. Ph. D. Dissertation. University of Plymouth.

[30] Mina Marmpena, Angelica Lim, Torbjørn S Dahl, and Nikolas Hemion. 2019. Generating robotic emotional body language with variational autoencoders. In *2019 8th International Conference on Affective Computing and Intelligent Interaction*. 545–551. https://doi.org/10.1109/ACII.2019.8925459

[31] Daniel Octavian Melinte and Luige Vladareanu. 2020. Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified Adam optimizer. *Sensors* 20, 8 (2020). https://doi.org/10.3390/s20082393

[32] Andre T. Nguyen, Luke E. Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. 2021. Practical cross-modal manifold alignment for robotic grounded language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1613–1622.

[33] Behnaz Nojavanasghari, Yuchi Huang, and Saad Khan. 2018. Interactive generative adversarial networks for facial expression generation in dyadic interactions. arXiv:1801.09092 [cs.CV]

[34] Ishaan Pakrasi, Novoneel Chakraborty, and Amy LaViers. 2018. A design methodology for abstracting character archetypes onto robotic systems. In *Proceedings of the 5th International Conference on Movement and Computing*. 1–8.

[35] Mannes Poel, Dirk Heylen, Anton Nijholt, M Meulemans, and A Van Breemen. 2009. Gaze behaviour, believability, likability and the iCat. *AI & Society* 24, 1 (2009), 61–73. https://doi.org/10.1007/s00146-009-0198-1

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*. 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[37] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161.

[38] Anvita Saxena, Ashish Khanna, and Deepak Gupta. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems* 2, 1 (2020), 53–79.

[39] Michael Suguitan, Randy Gomez, and Guy Hoffman. 2020. MoveAE: Modifying affective robot movements using classifying variational autoencoders. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 481–489. https://doi.org/10.1145/3319502.3374807

[40] Michael Suguitan and Guy Hoffman. 2019. Blossom: A handcrafted open-source robot. *ACM Transactions on Human-Robot Interaction* 8, 1, Article 2 (2019), 2:1–2:27 pages. https://doi.org/10.1145/3310356

[41] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics* 39, 6, Article 222 (2020). https://doi.org/10.1145/3414685.3417838

[42] Abylay Zhumekenov, Malika Uteuliyeva, Olzhas Kabdolov, Rustem Takhanov, Zhenisbek Assylbekov, and Alejandro J. Castro. 2020. Fourier neural networks: A comparative study. *Intelligent Data Analysis* 24 (2020). Issue 501. https://doi.org/10.3233/IDA-195050