MDPI

*Article*

# Emotion Transfer for 3D Hand and Full Body Motion using StarGAN

**Jacky C. P. Chan [1] and Edmond S. L. Ho [2]***

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong; cpchan@comp.hkbu.edu.hk
[2] Department of Computer and Information Sciences, Northumbria University, United Kingdom; e.ho@northumbria.ac.uk
* Correspondence: e.ho@northumbria.ac.uk;

**Abstract:** In this paper, we propose a new data-driven framework for 3D hand and full-body motion emotion transfer. Specifically, we formulate the motion synthesis task as an image-to-image translation problem. By presenting a motion sequence as an image representation, the emotion can be transferred by our framework using StarGAN. To evaluate our proposed method's effectiveness, we first conducted a user study to validate the perceived emotion from the captured and synthesized hand motions. We further evaluate the synthesized hand and full body motions qualitatively and quantitatively. Experimental results show that our synthesized motions are comparable to the captured motions and those created by an existing method in terms of naturalness and visual quality.

**Keywords:** hand animation; body motion; skeletal motion; emotion; motion capture; generative adversarial network; style transfer; user study

**Citation:** Chan, J. C. P.; Ho, E. S. L. Emotion Transfer for 3D Hand and Full Body Motion using StarGAN. *Computers* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:



## 1. Introduction

Effectively expressing emotion is crucial to improve the realism of 3D character animation. While animating facial expressions to reflect the character's emotional states has been an active research area [1–4], less attention has been paid to expressing emotion by other body parts, practically 'the body language'. In this work, we propose a general framework for synthesizing new hand and full body motions from an input motion, namely *emotion transfer*, by specifying the target emotion label. Our objective is to create motions for the character to present four emotions: anger, sadness, fear, and joy. Building upon our pilot study [5], we found that hand motion plays a vital role in computer animation since the subtle hand gestures can express a lot of different meanings and are useful for understanding a person's personality [6]. A classic example would be the character *Thing T. Thing* of the *"The Addams Family"* which is a hand, and it can 'act' and express a lot of different emotions solely by the fingers and hand movements. It is not surprising to see researchers proposing frameworks [7,8] for synthesizing hand and finger movements based on the given full-body motion to improve the expressiveness of the animation.

However, synthesizing hand motion is not a trivial task. Capturing hand motion using an optical motion capture system is not easy as the fingers are in proximity, and the labeling of the markers can be mixed up easily. As a result, most of the previous hand motion synthesis frameworks are based on physics-based motion generation models [8–12]. Recently, more effective hand motion capturing approaches are proposed. Alexanderson et al. introduce a new system for a passive optical motion capture system that can better obtain correct markers labels of fingers in real-time [13]. Han et al. [14] improve the difficulties in marker labeling for the optical MOCAP system using convolutional neural networks. While hand motion can be synthesized or captured using the approaches mentioned above, those motions are always challenging to be

reused because of the difficulties in transferring the styles to improve the expressiveness in different scenes.

This paper focuses on validating the effectiveness of the StarGAN-based emotion transfer framework proposed in our pilot study [5], which consist of three main components as illustrated in Figure 1: 1) converting motion data into image representation, 2) synthesizing new image by specifying the target emotion label using StarGAN, and 3) converting the synthesized image into motion data for animating 3D mesh models. In particular, we conducted a user study on evaluating how users perceived the emotion from the hand motions captured in [5] and those synthesized by our method. The naturalness and visual quality of the motions synthesized by our method are also evaluated and compared with exiting work [15]. We further demonstrate the generality of the proposed framework by transferring emotion on full-body 3D skeletal motions.

The contributions in this work can be summarized as follows:

- We proposed a new framework for transferring emotions in synthesizing hand and full body skeletal motions, which is built upon the success of our pilot study [5].
- We conducted a user study to validate the perceived emotion on the dataset we captured and open-sourced in our pilot study [5].
- We provide qualitative and quantitative results on the hand and full-body motion emotion transfer using the proposed framework, showing its validity by comparing them to captured motions.

## 2. Related Work

### *2.1. Hand animation*

Examples of hand animation can be easily found in various applications such as movies, games, and animations. However, capturing hand motion using existing motion capture systems is not a trivial task. There was no commercial or academic real-time vision-based hand motion capture solution until a recent work presented by Han et al. [14]. As a result, most of the previous work focused on synthesizing hand motions based on physics-based models. Liu [9] proposed an optimization-based approach for synthesizing hand-object manipulations. Given the initial hand pose for the desired initial contact, the properties of the object to interact and the kinematics goals, the 2-stage physics-based framework will synthesize the reaching and object manipulation motions accordingly. Andrews and Kry [10] proposed a hand motion synthesis framework for object manipulation. The method divides a manipulation task into 3 phases (approach, actuate, and release), and each phase is associated with a control policy for generating physics-based hand motion.

Liu et al. [11] introduced an optimization-based approach to hand manipulation of grasping pose. A physically plausible hand animation will be created by providing the grasping pose and the partial trajectory of the object. Ye and Liu [8] proposed a physics-based hand motion synthesis framework to generate detailed hand-object manipulations which match seamlessly with the full-body motion with wrist movements provided as the input. With the initial hand pose driven by the wrist movements, feasible contact trajectories (i.e. contacts between the hand/fingers and the object) will be found by random sampling at every time-step. Finally, detailed hand motion will be computed by using the contact trajectories as constraints in spacetime optimization. Bai and Liu [12] presented a solution to manipulate the orientation of a polygonal object using both the palm and fingers of a robotic hand. Their method considers the physical properties such as collisions, gravitational, and contact forces.

A PCA-based framework is proposed for data-driven approaches in [16] for generating detailed hand animation from a set of sparse markers. Jörg et al [7] proposed a data-driven framework for synthesizing the finger movements for an input full-body motion. The methods employ a simple approach for searching for an appropriate finger motion from the pre-recorded motion database based on the input wrist and body motion. While a wide range of approaches for hand motion synthesis are presented in the

literature, less attention has been paid to synthesizing expressive hand motions using high-level and intuitive control. Irimia et al. [15] proposed a framework for generating hand motion with different emotion by interpolation in the latent space. For every hand motion captured with different emotions, the hand poses are collected and projected to the latent space using PCA. New motion can be created by interpolating the hand poses using the latent representation. In contrast, our framework enables emotion transfer between different hand motions.

### *2.2. Style transfer for motion*

Motion style transfer is a technique used to convert the style of a motion to another style, thus creating new motions without losing the primitive content of the original one. An early work by Unuma et al. [17] proposed using Fourier principles to create an interactive and real-time control of locomotion with emotion, and include cartoon-ish exaggerations and expressions. Amaya et al. [18] introduced a model that could emulate emotional animation using signal processing techniques. The emotional transform is based on the speed and spatial amplitude of the movements.

Brand et al. [19] proposed a learned statistical model to synthesize styled motion sequences by interpolating and extrapolating the motion data. Urtasun et al. [20] lowered the dimension of the motion data by principal component analysis (PCA) and model the style transfer as the difference between the features. Ikemoto et al. [21] edited the motion using Gaussian process models of dynamics and kinematics for motion style transfer. Xia et al. [22] proposed a time-varying mixture of autoregressive models to represent the style difference between two motions. Their method learns such models automatically from unlabeled heterogeneous motion data.

Hsu et al. [23] presented a solution for translating the style of a human motion by comparing the difference of the behaviour of the aligned input and output motions using a linear time-invariant model. Shapiro et al. introduced a novel method of interactive motion data editing based on motion decomposition, which separates the style and expressiveness from the main motion [24]. The method uses Independent Component Analysis (ICA) to separate the style from the motion data.

Machine learning is applied to learn the style transfer from samples. Holden et al. leveraged a Convolution neural network to learn the style transfer from unaligned motion clips [25]. Smith et al. proposed a compact network architecture for learning the style transfer, which focuses on pose, foot contact, and timing [26]. To learn a motion controller with behavior styles applicable to unseen environment, Lee and Popović proposed an inverse reinforcement learning-based approach that works with a small set of motion samples [27].

Until now, the only research of style transfer was for a full-body character. This paper will propose a general method to full body character and the human hand. While we share a similar interest with the pilot study [15] on synthesizing hand motion with emotion, the previous work is technically interpolating emotion strength instead of emotion transfer.

#### 2.2.1. Image style transfer

Inspired by the encouraging results in image style transfer, we proposed formulating the emotion transfer for motion synthesis as an image-to-image translation problem. In this section, we review the recently proposed approaches in image style transfer.

Selim et al. [28] presented a new technique of style transfer that uses Convolutional Neural Networks (CNN) for extracting features from the input images. The method uses style transfer to transfer the features from a portrait painting to a portrait image. The method is generic and different kinds of styles can be transferred given the training data contains the required styles. To maintain the integrity of the facial structure and to capture the texture from the painting, the method uses spatial constraints such as the local transfer of the colour distribution. Elad et al. [29] presented a method for

transferring the style from painting to image indifferently of the portrayed subject. The method uses a fast style transfer process that gradually changes the resolution of the output image. To obtain the result, the creators applied multi-patch sizes and different resolution scales of the input source. The method is also able to control the colour pallet for the output image depending on the desire of the developer. Matsuo et al. [30] presented another style transfer method that uses CNN by combining a neural style transfer method with segmentation to obtain a partial texture style transfer. The method uses a CNN-based weakly supervised semantic segmentation technique and transfers the style to selected areas of the picture while maintaining the image's structure. The method uses neural style transfer to change the style of the selected part of the image. Unfortunately, a problem appears when the sources fail to map the style transfer, changing the background of the image even when the user does not select that area. In this work, we will represent the motion features as an image and CNN will be used in the core network.

## 3. Methodology

In this section, the proposed emotion transfer framework will be presented, and the overview is shown in Figure 1. Firstly, we introduce two datasets, hand motion database captured using Senso VR Glove and a full-body motion database captured using the MOCAP system (Section 3.1). These two databases contain motions with various emotional states and types. Next, the captured motions are standardized (Section 3.2) as a pre-processing step for the learning process. The motion data will then be transformed into an RGB image representation for learning the emotion transfer model using StarGAN. The StarGAN model learns how to generate a new image given a target domain label and the input image (Section 3.3). Finally, the synthesized new image will be converted to the joint angle/position space for generating the final 3D animation (Section 3.3.3). The details of each step will be explained in the following subsections.

### *3.1. Motion Datasets*

To learn how the motion features are mapped to emotion status, motion data is collected for training the models to be proposed in this paper. In particular, we used two datasets, which include the hand motion dataset collected in our pilot study [5] for hand motion synthesis, and the 3D skeletal motions in Body Movement Library [31] for full-body motion synthesis.

#### 3.1.1. Hand Motion with Emotion Label

We start with the details of hand motion dataset which was captured in our pilot study [5]. High-quality 3D skeletal hand motions were captured using the Senso VR glove (https://senso.me/). There are 35 motions in total, with 7 different action types, including *Crawling*, *Gripping*, *Patting*, *Impatient*, *Hand on Mouse*, *Pointing*, and *Pushing*. Each motion types are captured under 5 different types of emotions and their characteristics are listed on Table 1. Readers are referred to [5] for the details of the data capturing process.

In this dataset, each hand motion at each frame is represented by a vector $P_j$

$$P_j = [p^0_{j,x}, p^0_{j,y}, p^0_{j,z}, ..., p^{n-1}_{j,x}, p^{n-1}_{j,y}, p^{n-1}_{j,z}] \quad (1)$$

where $j$ is the index of the frame, $n$ is the joint number in the 3D hand skeletal structure and $n = 27$ in all of the data we capture, and $p$ contains the joint rotations on the x,y and z axes, respectively. Therefore, each keyframe is a 81-dimensional feature vector. The hand translation was discarded as in [5,15] due to the inconsistent global locations of the hand in the captured motions.

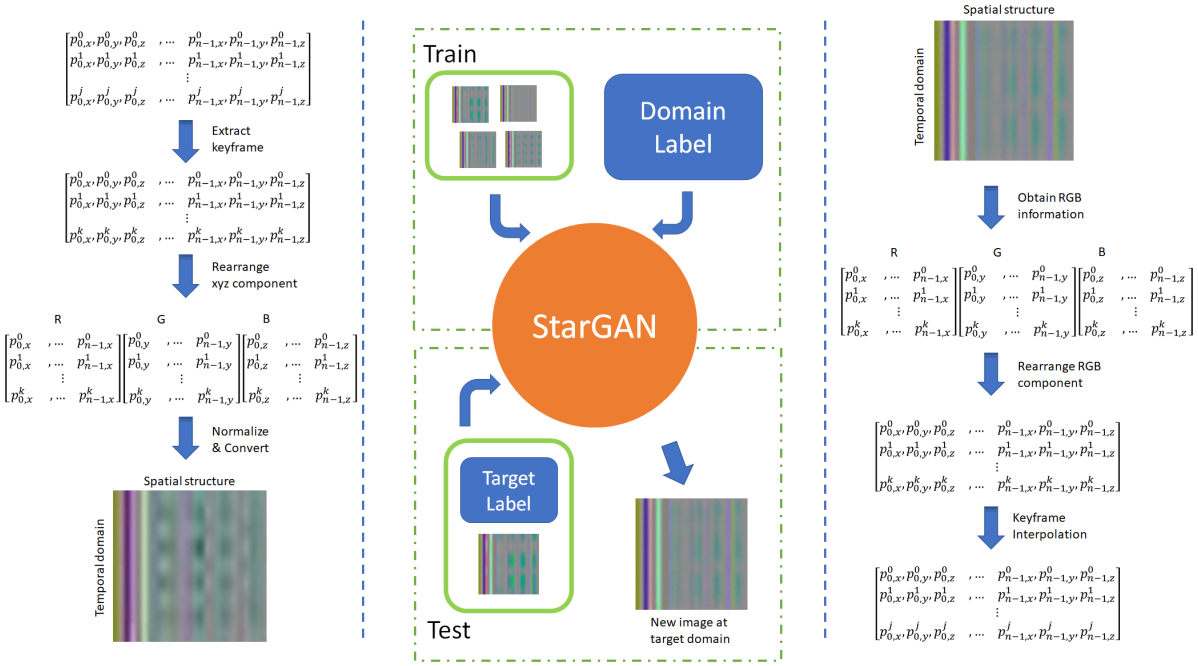


**Figure 1.** The overview of the proposed emotion transfer framework. (left) Convert motion data to image format. (middle) StarGAN learn how to generate realistic fake image given a sample and target domain label. (right) Obtain new motion data from the generated image

| Emotion | Characteristics |
|---|---|
| Angry | exaggerated, fast, large range of motion |
| Happy | energetic, large range of motion |
| Neutral | normal, styleless motion |
| Sad | sign of tiredness, small range of motion |
| Fearful | asynchronous finger movements, small range of motion |

Table 1: The 5 types of emotions used in the hand motion dataset [5] and their characteristics.

#### 3.1.2. Full Body Motion with Emotion Label

Here, the details of the full-body motion dataset are presented. The Body Movement Library [31] were captured with the Falcon Analog optical motion capture system. In order to capture the emotion expressions naturally from the subjects, 30 nonprofessional participants (15 females and 15 males) with an age range from 17 to 29 years old were recruited. Three motion types, *knocking, lifting, and throwing*, are included in our experiment. A skeletal structure with 33 joints was used in all of the captured motions. Note that only the 3D joint positions in Cartesian coordinates are available. There are 3 motion types in the dataset, including *knocking, lifting and throwing*. Each subject performed each motion type with four different emotion status: *Neutral, Angry, Happy, and Sad*. The dataset contains 4,080 motions in total.

### *3.2. Standardizing Motion Feature*

Due to the environmental setting and personal style, the captured hand motion data varies significantly representation both spatially and temporally. Data standardization (or normalization) is used to facilitate the learning process in the later stage. While some advanced techniques such as Recurrent Neural Network (RNN) can be used to model

data sequences with variations in length, such a method requires a significant amount of data to train the model, which is not feasible with the dataset that have been collected.

To handle the temporal difference, keyframes are extracted from the motion by curve simplification to facilitate the learning process. By considering the reconstruction errors when interpolating the in-between motion using spline interpolation, we found the optimal numbers of keyframes for every type of motion. For hand motion, good performance can be achieved by extracting 9 keyframes as in [5].

As a result, each hand motion sequence is represented by a vector $M$ in the joint angle space:

$$M = [Pk_0, ..., Pk_{k-1}, Pk_k] \tag{2}$$

where $k$ is the total number of keyframes and $k = 9$, $Pk_i$ is the $i$-th keyframe and $Pk$ is having the same representation as in $P$ (Eq. 1).

For full-body motions, we empirically found that the optimal number of keyframes of knocking, throwing, lifting are 13, 20, and 13, respectively. In [32], Chan et al. observed that people express different emotions by using different speeds and rhythms. Such an observation aligns well with the characteristics we found in hand motions as presented in Table 1. Specifically, there is a significant difference in the speed of body movements. For example, the arm of the subject swings faster in the *happy* throwing motion than the *sad* one (see Figure 8). To better represent this key characteristic, we compute the joint velocity between adjacent keyframes as follows:

$$v_{i+1} = (k_{i+1} - k_i)/\Delta t \tag{3}$$

where $\Delta t$ is the duration between the two adjacent keyframes. Therefore, each keyframe is represented by 3D joint positions and velocities and results in a 99 + 99 = 198-dimensional feature vector. The full body motion sequence can then be represented as a sequence of keyframes as in Eq. 1.

### *3.3. Emotion transfer*

Generative adversarial network (GAN) based framework gain much attention in the area of Computer Graphics. Encouraging results are found in style transfer frameworks such as CycleGAN [33] and DualGAN [34] for image-to-image translation. These results inspire us to adopt such kind of framework for emotion transfer for skeletal motion synthesis tasks.

In the rest of this section, we will first explain how to represent a motion in the format of an image in Section 3.3.1. Next, the justifications on adopting the StarGAN [35] framework will be given in Section 3.3.2. Finally, a new motion will be reconstructed from the synthesized image (Section 3.3.3).

#### 3.3.1. Representing motion as an image

In order to use the Image-to-Image domain translation framework for motion emotion transfer, we will show how to represent a motion sequence as an image. The x, y and z components (i.e. angles for hand motions and positions for full-body motions) are arranged chronologically as the RGB components of the image. Each frame of a motion is represented as a row of an image, while each joint of a motion is represented as a column of an image. Hence, each keyframed hand motion $M$ will be arranged as a 3-channel $(27 \times k)$ matrix. On the other hand, each keyframed full-body motion $M_{full}$ will be arranged as a 3-channel $(66 \times k)$ matrix. The values in the 3-channel matrix are then re-scaled into the range of [0, 255], which is the typical range of RGB value, as follows:

$$v^m_{i,c} = round(255 \times \frac{(p^m_{i,c} - p_{min})}{(p_{max} - p_{min})}) \tag{4}$$

where $m$ is the joint index, $i$ is the keyframe index, $c \in \{x, y, z\}$ represents the channel index, $V^m_{i,c}$ is the normalized pixel value, $p_{max}$ and $p_{min}$ are the maximum and minimum

values among all the joint angles/positions/velocities existed in the dataset. Noted that the images are saved in Bitmap format to avoid data loss during compression. Examples of the image representation of the motions are illustrated in Figure 2 and 3. It can be seen that the different motions are represented by different image patterns, which will be useful for extracting discriminative patterns in the learning process. From Figure 3, we can see that the main difference between the four emotions is at the right-hand side of the images, which is about joint velocity.

When converting the motion into an image, sequential ordering is used. Such an approach is commonly used for arranging the data of each joint [36]. In this study, our main focus is to evaluate the performance of adapting the StarGAN network for emotion transfer in motion synthesis tasks. As a result, we directly utilize the original StarGAN network architecture which contains 2D Convolutional layers. Since 2D Convolution focuses on local neighbours (i.e. image pixels nearby) only, using the sequential ordering method can result in sub-optimal results when representing the tree-like skeletal structure for full-body and hand motions as the neighbouring joints are not necessarily close-by after converting into the image representation. While encouraging results are obtained in this study, we will explore the use of other approaches to better represent motions in the StarGAN framework in the future, such as Graph Convolutional Networks (GCNs) [37] and its variants [38] which demonstrated a better performance in modelling human-like skeletal motions.

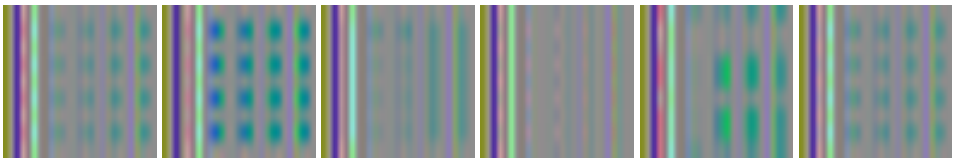

**Figure 2.** Examples of the image representation of neutral hand motions. From left to right: Crawling, Gripping, Impatient, Patting, Pointing and Pushing.

**Figure 3.** Examples of the image representation of full body motions (throwing) under different emotions. From left to right: Angry, Happy, Neutral and Sad.

#### 3.3.2. Emotion transfer as Image-to-Image domain translation

One of the potential applications of the proposed system is to create new motion by controlling the *emotion labels*. To support the translation between multi-domain and considering the robustness and scalability, StarGAN [35] is adapted to translate motion from one emotion to another emotion while preserving the basic information of the input motion. Comparing to typical GANs with cycle consistency losses such as CycleGAN [33] and DualGAN [34] for style transfer, StarGAN [35] can perform image-to-image translations for multiple domains using only a single model, which is suitable for transferring different types of emotions. Readers are refereed to our pilot study [5] and [35] for the technical details.

#### 3.3.3. Reconstructing Hand Motion from Generated Images

Since the output of StarGAN is an image, we need to reconstruct it to obtain the new motion (i.e., joint position/angle space). The first step is to re-scale the RGB values:

$$p^m_{i,c} = (\frac{v^m_{i,c}}{255} \times (p_{max} - p_{min})) + p_{min} \quad (5)$$

where $v_{i,j}^{m}$ is the pixel value for the $c$-th channel of the $m$-th row at $i$-th column on the synthesized image, $p_{max}$ and $p_{min}$ are same values as in Equation 4. Next, we rearrange the pixel values to convert the image representation back to the keyframed motion $M$. Then the duration between two adjacent keyframes can be approximated by using joint velocity:

$$\Delta t = (k_{i+1} - k_i)/v_{i+1} \tag{6}$$

Finally, the new motion is produced by applying spline interpolation on those keyframes.

## 4. Experimental Results

To evaluate the effectiveness of the proposed emotion transfer framework, a wide range of experiments are conducted to assess the performance qualitatively and quantitatively. In particular, we first carried out a user study (Section 4.1) to understand how users perceive the emotions from the captured and synthesized hand motions. Next, a series of hand and full body animations are synthesized to compare different emotions visually (Section 4.2). To demonstrate the framework's practicality, we employ a leave-one-out cross-validation approach to split the data into training, and testing sets. The results presented in the section are synthesized from unseen samples. We employed a leave-one-out cross-validation approach for hand motions as in the existing approaches [5,15]. For each action type, we only captured 1 sample for each emotion type in the hand motion dataset. As a result, the leave-one-out cross-validation is the best data-split approach we can use in order to 1) maximize the number of training data, while 2) keeping the testing sample to be 'unseen' during the training process. For full-body motions, the motions of 26 subjects are used for training, while the motions of the remaining 4 subjects are used for synthesizing the results. The readers are referred to the video demo submitted as supplementary materials for more results.

### *4.1. User Study on Hand Animations*

A user study was conducted to evaluate how users perceived the emotion from each hand animation created by either the captured motions or synthesized by our proposed method. The study was published online using Google Form, and the animations are embedded in the online form to facilitate the side-by-side comparison. We invited a group of final year undergraduate students who are studying in the Computer Science programme. In particular, the students completed a course on Computer Graphics and Animation recently, although the course does not cover any specific knowledge on hand animation and emotion-based motion synthesis. The range of age in the group is 20-25. At the end of the study, 30 completed sets of the survey were received. The study is divided into several parts, including the emotion recognition from the captured dataset (Section 4.1.1) and synthesized animations (Section 4.1.2) for validating whether the participants perceived the emotion correctly, and the evaluation of naturalness and visual quality of the captured motions, synthesized motions and results created by Irimia et al. [15] (Section 4.1.3). The details are explained in the following subsections.

#### 4.1.1. Evaluating the emotion perceived from the captured hand motions

To evaluate how users perceive emotion from hand animation, we first analyze the accuracy of emotion recognition from the captured hand motions. Specifically, we randomly select 4 hand animations from the capture motions, which include one motion in each emotion category (i.e. *Angry, Happy, Sad, and Afraid*). Note that no briefing, such as the characteristics for each emotion class as listed on Table 1, is provided to the users. As a result, the users labeled each hand animation based on their interpretation of the emotions. For each animation, the *neutral* emotion of the hand animation is shown to the user first. Next, the user was asked to choose the most suitable emotion label for the hand animation with emotion. The averaged and class-level emotion recognition accuracies are reported in Table 2 (see the 'Captured' column).

| Emotion | Captured | Synthesized |
|---|---|---|
| Angry | 70.00% | 73.33% |
| Happy | 60.00% | 60.00% |
| Sad | 70.00% | 60.00% |
| Fearful | 63.33% | 70.00% |
| Average | 65.83% | 65.83% |

Table 2: Emotion recognition accuracy (%) on the captured and synthesized hand motions in the user study.

| | | Perceived Emotion | | | |
|---|---|---|---|---|---|
| | | Angry | Happy | Sad | Feared |
| Ground Truth Labels | Angry | **70.00%** | 15.00% | 3.33% | 11.67% |
| | Happy | 23.33% | **60.00%** | 6.67% | 10.00% |
| | Sad | 6.67% | 10.00% | **70.00%** | 13.33% |
| | Feared | 13.33% | 10.00% | 13.33% | **63.33%** |

Table 3: Confusion matrix of the emotion recognition test on the captured hand motions in the user study.

The averaged recognition accuracy is 65.83%, which shows that the majority of the users perceived the correct emotion from the captured hand motions. The class-level recognition accuracies further show the consistency among all classes. In particular, a good recognition accuracy at 70.00% was achieved in both Angry and Sad classes. To further analyze the results, the confusion matrix of the emotion recognition test is presented in Table 3. It can be seen that the recognition accuracy of the Happy class is lower than other classes. This is mainly caused by the inter-class similarity between happy and angry since the motions in both classes are having a large range of motion. Although ambiguity can be found between motions from different classes, the results highlight the correct recognition is dominating the results.

#### 4.1.2. Evaluating the emotion perceived from the hand motions synthesized by our method

Similar to Section 4.1.1, we also analyze the accuracy of emotion recognition from the hand motions synthesized using our method. Again, each user was asked to label 4 synthesized hand animations. The averaged and class-level emotion recognition accuracies are reported in Table 2 (see the 'Synthesized' column). The averaged recognition accuracy is 65.83%, which is the same as the accuracy obtained from the captured dataset. This result highlights no significant difference in the expressiveness of emotion between the captured and synthesized hand motions. While the averaged recognition accuracies are the same, it can be seen that the class-level accuracies are different, and the confusion matrix is presented in Table 4. The results indicate that the synthesized motions demonstrated a lower level of inter-class similarity. The perceived emotion only spread across 3 classes for Angry, Sad and Feared (i.e. with one class having 0% of voting) instead of 4 classes as in the results obtained from captured motions.

While the inter-class similarity is reduced in general, it can be observed that some pairs of classes are having higher similarity in the motions synthesized by our method. For example, the 'Sad-Happy' pair (i.e. misperceiving *Sad* as *Happy*) increases from 10% to 21.67%. this can be caused by the small changes in the speed of the motions as the synthesized motions tend to have a slightly smaller range of speed difference across different emotions. This affects the *Sad* and *Happy* motions since the major difference between these 2 classes is the speed. The 'Feared-Sad' and 'Sad-Feared' also demonstrated an increase in ambiguity. This can be caused by the averaging effect by training the StarGAN model using all different types of motions. The most discriminative characteristics of the *Feared* class is the asynchronous finger movement.

| | | Perceived Emotion | | | |
|---|---|---|---|---|---|
| | | Angry | Happy | Sad | Feared |
| Ground Truth Labels | Angry | **73.33%** | 16.67% | 0.00% | 10.00% |
| | Happy | 23.33% | **60.00%** | 6.67% | 10.00% |
| | Sad | 0.00% | 21.67% | **60.00%** | 18.33% |
| | Feared | 6.67% | 0.00% | 23.33% | **70.00%** |

Table 4: Confusion matrix of the emotion recognition test on the synthesized hand motions in the user study.

| | |
|---|---|
| Synthesized is more pleasant | 46.67% |
| Do not know | 13.33% |
| Captured is more pleasant | 40.00% |

Table 5: Emotion recognition accuracy (%) on the captured and synthesized hand motions in the user study.

Learning a generic model for all different emotions and action types has reduced the discriminative characteristics in the synthesized motions. It will be an interesting future direction to explore a better way to separate the emotion, action and personal style into different components in the motion to better preserve the motion characteristics.

#### 4.1.3. Comparing the naturalness and visual quality of the synthesized animations with the captured motions and baseline

To evaluate the visual quality and naturalness of the synthesized hand animations, we follow the design of the user study conducted in [39] by playing back the captured and synthesized motions side-by-side, and ask the user which one is 'more pleasant' or the user can answer 'do not know' if it is difficult to judge. In this experiment, 4 pairs of animations were randomly selected from our database and shown to each user. We randomly selected *non-neutral* motions from the captured data and paired them up with the corresponding synthesized motions, which are emotionally transferred from *neutral* to the target emotion state. The results are summarized in Table 5. From the results, it can be seen that both of the synthesized and captured motions have received a similar percentage of users rating as 'more pleasant', while the motion produced by our method is 6.67% higher than the captured motions. There are 13.3% of users who cannot decide which motion is better. To validate the user study results, A/B testing is used to find out the statistical significance of the results. By treating the captured motions as variant A and our synthesized motions as variant B, the p-value is 0.2301. This suggests the conclusion on *'the synthesized motions are visually more pleasant than the captured motions'* is not statistically significant at the 95% confidence interval. On the other hand, when including our synthesized motions and the 'do not know' option in variant B, the p-value becomes 0.0127. This suggests the conclusion on *'the synthesized motions are not visually less pleasant than the captured motions'* is statistically significant at the 95% confidence interval. In summary, the visual quality between the captured and synthesized hand motions are similar, and arguably our method will not degrade the visual quality of the input hand motion, as the results indicate.

We further compared the motions synthesized by our method with those created using a PCA-based method proposed by Irimia et al. [15]. Again, we follow the user study explained above to evaluate the difference in the visual quality and naturalness of the synthesized motions. A side-by-side comparison will be given to the user to rate whether ours or Irimia et al. [15]'s method produces 'more pleasant' animation or no decision can be made. Each user was asked to rate 4 pairs of animations, and the results are presented in Table 6. The results show that the animations synthesized by our methods have better visual quality than those generated by Irimia et al. [15] with a more significant margin of 8.33% more users rated our results as 'more pleasant', although

| | |
|---|---|
| Synthesized is more pleasant | 45.00% |
| Do not know | 18.33% |
| Irimia et al. [15] is more pleasant | 36.67% |

Table 6: Emotion recognition accuracy (%) on the captured and synthesized hand motions in the user study.

the p-value ($p = 0.1757$) computed in the A/B test suggests that the results are not statistically significant at the 95% confidence interval. Similar to the comparison between the captured and synthesized motions, we group our synthesized motions and the 'do not know' option and compare with the results created generated by Irimia et al. [15]. The p-value becomes 0.0012 which suggests *'our synthesized motions are not visually less pleasant than those generated by Irimia et al. [15]'*. Again, the results highlight the methods compared in this study are producing motions with similar visual quality. In addition to the visual quality, the capability of multi-class emotion transfers in our proposed method is another advantage over Irimia et al. [15] since their approach is essentially interpolating the motion between two different emotional states.

### *4.2. Evaluation on Emotion Transfer*

In this section, a wide range of results synthesized by our method are presented. We will first present the synthesized hand animations in Section 4.2.1 and 4.2.2. Next, the synthesized full-body animations will be discussed in Section 4.2.3. The animations are also included in the accompanying video demo.

#### 4.2.1. Hand Animation

Here, we demonstrate the effectiveness of the proposed method by showing some of the synthesized hand animations. Like the experiments mentioned earlier, we used an unseen neutral hand motion as input and synthesized the animations by specifying the emotion labels. Due to the limited space, we visualize the results (Figure 4a to 4d) on 4 motion sets including *crawling*, *patting*, *impatient* and *pushing*. In each figure, each row contains 5 hand models which are animated by motions with different emotions (from left to right): *angry*, *happy*, *input (neutral)*, *sad*, *fearful*. The 4 rows in each figure are referring to the keyframes of the 4 progression stages (i.e. 0%, 33%, 67%, and 100%) in each animation.

The experimental results are consistent. To assess the correctness of the synthesized motion, we can compare the changes of the motion between keyframes in each column in each figure and evaluate if the changes align with the characteristics listed in Table 1. Specifically, input motions become more exaggerated by transferring to *angry*. The range of motion increases, and the motion becomes faster. This is highlighted by the movement of the thumb in the video demo. By transferring to the *happy* emotion, the motion becomes more energetic with a larger range of motion when compared with the input (*neutral*) motion. The motion's speed is getting higher as well, although the motion is less exaggerated than those transferred to *angry*. With the *sad* emotion, the synthesized motions show the sign of tiredness, which results in slower movement. Finally, the *fearful* emotion brings the asynchronous finger movements to the neutral motion as those characteristics can be found in the captured data. In summary, the consistent observation of the synthesized motions highlighted the effectiveness of our framework.

#### 4.2.2. Comparing Emotion Transferred Motions with Captured Data

Next, we compare the synthesized motions with the captured data. Recall that leave-one-out cross-validation is used in defining the training and testing data sets. As a result, the motion type of the input (i.e., testing) motion is not included in the training data. It is possible that the action of the synthetic motion looks slightly different from

the captured motion. Having said that, an effective emotion transfer framework should be able to transfer the characteristics of the corresponding emotion to the new motion.

The results are illustrated in Figure 5a to 5e. Similar to Section 4.2.1, we extracted 3 keyframes (i.e. each row in each figure) at the different progression stages (0%, 50% and 100%) of the animation. The hand models in each column (from left to right) were animated by the input (*neutral*), synthesized (i.e. emotion transferred) and captured motions, respectively. Here, readers can focus on whether the synthesized motions (middle column in each figure) contains the characteristics of the corresponding emotion as in the captured motions (right column in each figure). Readers can also compare the difference between the input (*neutral*, left column in each figure) and the synthesized motion to evaluate the changes made by the proposed framework.

It can be seen that the motions synthesized by the proposed framework have the characteristics of the corresponding emotion. For example, the motions are exaggerated in the *angry* crawling and pushing motions in Figure 5a and 5d, respectively. On the other hand, the *sad* crawling motion shows the sign of tiredness. The *fearful* emotion can again transfer the asynchronous finger movements to the pushing motion, as illustrated in Figure 5b. Finally, a larger range of motion can be seen in the *happy* impatient motion (Figure 5c).

#### 4.2.3. Body motion synthesis results

To further demonstrate the generality of the proposed framework, we trained the proposed framework using 3D skeletal full-body motion in this experiment. Again, unseen motions with the *neutral* emotion are used as input, and new motions are synthesized by specifying the emotion labels. Three types of motions, including knocking, lifting, and throwing, are included in the test and the screenshots of some examples are illustrated in Figure 6 to 8. We selected 3 key moments from each animation representing the early, middle, and late stages of the motion in the screenshots. To facilitate the side by side comparison, we show the input motion (blue), synthesized motions (green) and the captured motions (purple) in Figure 6 to 8.

In general, there is a consistent trend in terms of the difference in the speed between motions with different emotions. Specifically, the *angry* motions are the fastest, with *happy* motions are slightly slower, *neutral* motions are the third, and *sad* motions are the slowest in most of the samples. Such pattern can be seen in the motions synthesized by our method. Another observation from the results is the small difference between the synthesized motions and the captured ones (i.e., ground truth). We believe the small difference is mainly caused by the proposed method emphasized learning emotion transfer without explicitly modeling the personal style differences from motions performed by different subjects. As a result, small differences can be introduced when synthesizing new motions by our method. This is an interesting further direction to incorporate personal style in the motion modeling process to further strengthen the proposed method.

## 5. Conclusion and Discussions

In this paper, we propose a new framework for synthesizing motion by emotion transfer. We demonstrated the generality of the framework by modeling hand and full body motions in a wide range of experiments. A user study is conducted to verify the perceived emotion from the hand motions as well as evaluating the visual quality and naturalness of the animations. Experimental results show that our method can 1) generate different styles of motions according to the emotion type, 2) the characteristics in each emotion type can be transferred to new motions, and 3) achieving similar or better visual quality when comparing the hand motions synthesized by our method with those captured motions and created by [15].

In the future, we are interested in incorporating the personal style into the motion modeling framework. In addition to specifying the emotion labels to synthesize different

motions, de-tangling the 'base' motion and personal style can further increase the variations in the synthesized motions. Another further direction will be evaluating the feasibility of having multiple emotion labels with different levels of strengths for motion representation and synthesis. Such a direction is inspired by the emotion recognition results of the user study in which not all users are agreed on a single type of emotion being associated with each hand motions in the study. It is also an interesting future direction to quantitatively evaluate the results by comparing the differences between the synthesized and ground-truth motion numerically. To achieve this goal, more hand motions have to be captured. As in our pilot study, we have difficulties capturing the global translation and rotation in high quality. As a result, the global transformation is discarded, which limits the expression of emotion. One possible solution is to capture the hand motions using state-of-the-art MOCAP solutions such as [14].

## Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## References

1. Karras, T.; Aila, T.; Laine, S.; Herva, A.; Lehtinen, J. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Trans. Graph.* **2017**, *36*. doi: 10.1145/3072959.3073658.
2. Tinwell, A.; Grimshaw, M.; Nabi, D.A.; Williams, A. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior* **2011**, *27*, 741–749. Web 2.0 in Travel and Tourism: Empowering and Changing the Role of Travelers, doi:https://doi.org/10.1016/j.chb.2010.10.018.
3. Courgeon, M.; Clavel, C. MARC: a framework that features emotion models for facial animation during human–computer interaction. *Journal on Multimodal User Interfaces* **2013**, *7*, 311–319. doi:10.1007/s12193-013-0124-1.
4. Ruttkay, Z.; Noot, H.; Ten Hagen, P. Emotion Disc and Emotion Squares: Tools to Explore the Facial Expression Space. *Computer Graphics Forum* **2003**, *22*, 49–53, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8659.t01-1-00645]. doi: https://doi.org/10.1111/1467-8659.t01-1-00645.
5. Chan, J.C.P.; Irimia, A.S.; Ho, E.S.L. Emotion Transfer for 3D Hand Motion using StarGAN. Computer Graphics and Visual Computing (CGVC); Ritsos, P.D.; Xu, K., Eds. The Eurographics Association, 2020. doi:10.2312/cgvc.20201146.
6. Wang, Y.; Tree, J.E.F.; Walker, M.; Neff, M. Assessing the Impact of Hand Motion on Virtual Character Personality. *ACM Trans. Appl. Percept.* **2016**, *13*. doi:10.1145/2874357.
7. Jörg, S.; Hodgins, J.; Safonova, A. Data-Driven Finger Motion Synthesis for Gesturing Characters. *ACM Trans. Graph.* **2012**, *31*. doi:10.1145/2366145.2366208.
8. Ye, Y.; Liu, C.K. Synthesis of Detailed Hand Manipulations Using Contact Sampling. *ACM Trans. Graph.* **2012**, *31*, 41:1–41:10. doi:10.1145/2185520.2185537.
9. Liu, C.K. Synthesis of Interactive Hand Manipulation. Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation; Eurographics Association: Aire-la-Ville, Switzerland, Switzerland, 2008; SCA '08, pp. 163–171.
10. Andrews, S.; Kry, P.G. Policies for Goal Directed Multi-Finger Manipulation. VRIPHYS, 2012.
11. Liu, C.K. Dextrous Manipulation from a Grasping Pose. ACM SIGGRAPH 2009 Papers; ACM: New York, NY, USA, 2009; SIGGRAPH '09, pp. 59:1–59:6. doi:10.1145/1576246.1531365.
12. Bai, Y.; Liu, C.K. Dexterous manipulation using both palm and fingers. 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 1560–1565. doi: 10.1109/ICRA.2014.6907059.
13. Alexanderson, S.; O'Sullivan, C.; Beskow, J. Robust online motion capture labeling of finger markers. Proceedings of the 9th International Conference on Motion in Games. ACM, 2016, pp. 7–13.
14. Han, S.; Liu, B.; Wang, R.; Ye, Y.; Twigg, C.D.; Kin, K. Online Optical Marker-Based Hand Tracking with Deep Labels. *ACM Trans. Graph.* **2018**, *37*. doi:10.1145/3197517.3201399.

15. Irimia, A.S.; Chan, J.C.P.; Mistry, K.; Wei, W.; Ho, E.S.L. Emotion Transfer for Hand Animation. Motion, Interaction and Games; ACM: New York, NY, USA, 2019; MIG '19, pp. 41:1–41:2. doi:10.1145/3359566.3364692.
16. Wheatland, N.; Jörg, S.; Zordan, V. Automatic Hand-Over Animation Using Principle Component Analysis. Proceedings of Motion on Games; Association for Computing Machinery: New York, NY, USA, 2013; MIG '13, p. 197–202. doi:10.1145/2522628.2522656.
17. Unuma, M.; Anjyo, K.; Takeuchi, R. Fourier Principles for Emotion-based Human Figure Animation. Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques; ACM: New York, NY, USA, 1995; SIGGRAPH '95, pp. 91–96. doi:10.1145/218380.218419.
18. Amaya, K.; Bruderlin, A.; Calvert, T. Emotion from Motion. Proceedings of the Conference on Graphics Interface '96; Canadian Information Processing Society: Toronto, Ont., Canada, Canada, 1996; GI '96, pp. 222–229.
19. Brand, M.; Hertzmann, A. Style machines. Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000, pp. 183–192.
20. Urtasun, R.; Glardon, P.; Boulic, R.; Thalmann, D.; Fua, P. Style-based motion synthesis. Computer Graphics Forum. Wiley Online Library, 2004, Vol. 23, pp. 799–812.
21. Ikemoto, L.; Arikan, O.; Forsyth, D. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics (TOG)* **2009**, *28*, 1–12.
22. Xia, S.; Wang, C.; Chai, J.; Hodgins, J. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)* **2015**, *34*, 1–10.
23. Hsu, E.; Pulli, K.; Popović, J. Style Translation for Human Motion. ACM SIGGRAPH 2005 Papers; ACM: New York, NY, USA, 2005; SIGGRAPH '05, pp. 1082–1089. doi:10.1145/1186822.1073315.
24. Shapiro, A.; Cao, Y.; Faloutsos, P. Style Components. Proceedings of Graphics Interface 2006; Canadian Information Processing Society: Toronto, Ont., Canada, Canada, 2006; GI '06, pp. 33–39.
25. Holden, D.; Habibie, I.; Kusajima, I.; Komura, T. Fast Neural Style Transfer for Motion Data. *IEEE Computer Graphics and Applications* **2017**, *37*, 42–49. doi:10.1109/MCG.2017.3271464.
26. Smith, H.J.; Cao, C.; Neff, M.; Wang, Y. Efficient Neural Networks for Real-Time Motion Style Transfer. *Proc. ACM Comput. Graph. Interact. Tech.* **2019**, *2*. doi:10.1145/3340254.
27. Lee, S.J.; Popović, Z. Learning Behavior Styles with Inverse Reinforcement Learning. *ACM Trans. Graph.* **2010**, *29*. doi:10.1145/1778765.1778859.
28. Selim, A.; Elgharib, M.; Doyle, L. Painting Style Transfer for Head Portraits Using Convolutional Neural Networks. *ACM Trans. Graph.* **2016**, *35*, 129:1–129:18. doi:10.1145/2897824.2925968.
29. Elad, M.; Milanfar, P. Style Transfer Via Texture Synthesis. *IEEE Transactions on Image Processing* **2017**, *26*, 2338–2351. doi:10.1109/TIP.2017.2678168.
30. Matsuo, S.; Shimoda, W.; Yanai, K. Partial style transfer using weakly supervised semantic segmentation. 2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2017, pp. 267–272. doi:10.1109/ICMEW.2017.8026228.
31. Ma, Y.; Paterson, H.M.; Pollick, F.E. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods* **2006**, *38*, 134–141.
32. Chan, J.C.P.; Shum, H.P.H.; Wang, H.; Yi, L.; Wei, W.; Ho, E.S.L. A generic framework for editing and synthesizing multimodal data with relative emotion strength. *Computer Animation and Virtual Worlds* **2019**, *30*, e1871, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.1871]. e1871 cav.1871, doi:10.1002/cav.1871.
33. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.
34. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2868–2876.
35. Choi, Y.; Choi, M.; Kim, M.; Ha, J.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
36. Holden, D.; Saito, J.; Komura, T. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* **2016**, *35*. doi:10.1145/2897824.2925975.

37. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (ICLR), 2017.
38. Men, Q.; Ho, E.S.L.; Shum, H.P.H.; Leung, H. A Quadruple Diffusion Convolutional Recurrent Network for Human Motion Prediction. *IEEE Transactions on Circuits and Systems for Video Technology* **2020**, pp. 1–1. doi:10.1109/TCSVT.2020.3038145.
39. Aristidou, A.; Zeng, Q.; Stavrakis, E.; Yin, K.; Cohen-Or, D.; Chrysanthou, Y.; Chen, B. Emotion Control of Unstructured Dance Movements. Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation; Association for Computing Machinery: New York, NY, USA, 2017; SCA '17. doi:10.1145/3099564.3099566.

**(a)** *crawling*

**(b)** *impatient*

**(c)** *pushing*

**(d)** *patting*

**Figure 4.** Screenshots (one frame per row) of transferring the input (neutral) motion to different types of emotions. Columns from left to right: *angry*, *happy*, *input (neutral)*, *sad*, *fearful*.

**(a)** *crawling*, transferred to *angry*

**(b)** *pushing*, transferred to *fearful*

**(c)** *impatient*, transferred to *happy*

**(d)** *pushing*, transferred to *angry*

**(e)** *crawling*, transferred to *sad*

**Figure 5.** Screenshots (one frame per row) of the comparison between the input (*neutral*, left), synthesized (i.e. emotion transferred, middle) and captured (right) motions.

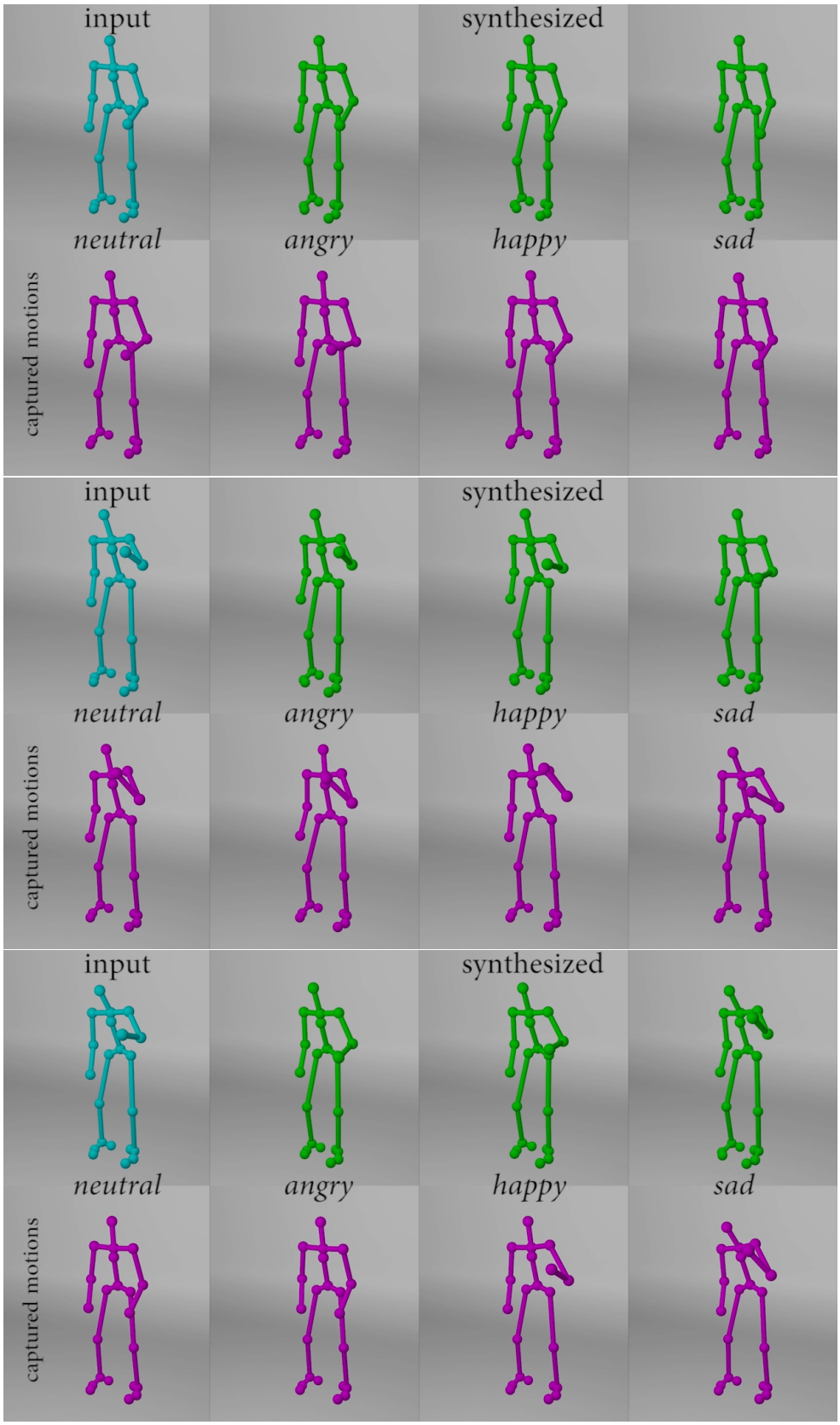


**Figure 6.** Screenshots of the *knocking* motion extracted from different stages - early (top row), middle (middle row), late (bottom row).

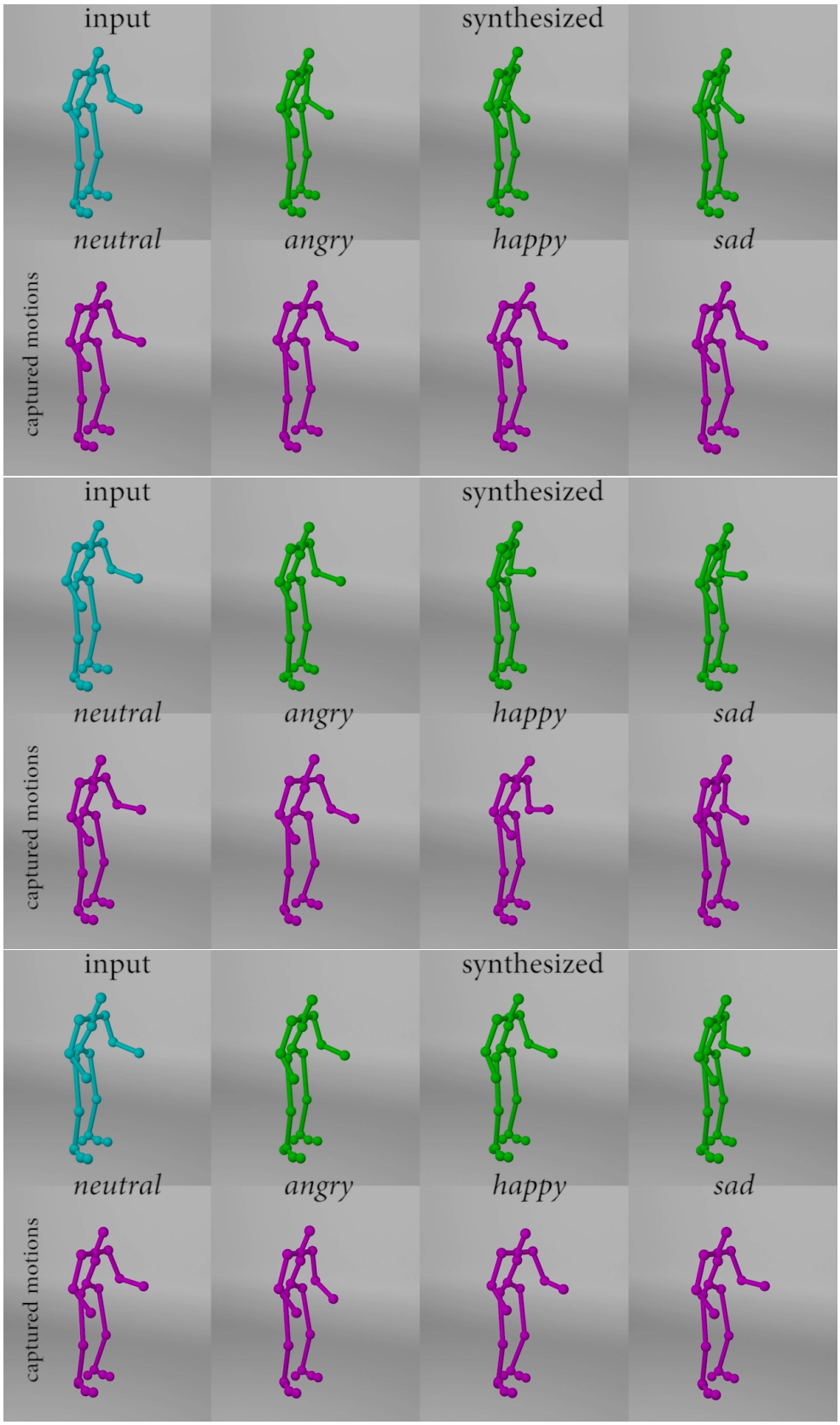


**Figure 7.** Screenshots of the *lifting* motion extracted from different stages - early (top row), middle (middle row), late (bottom row).

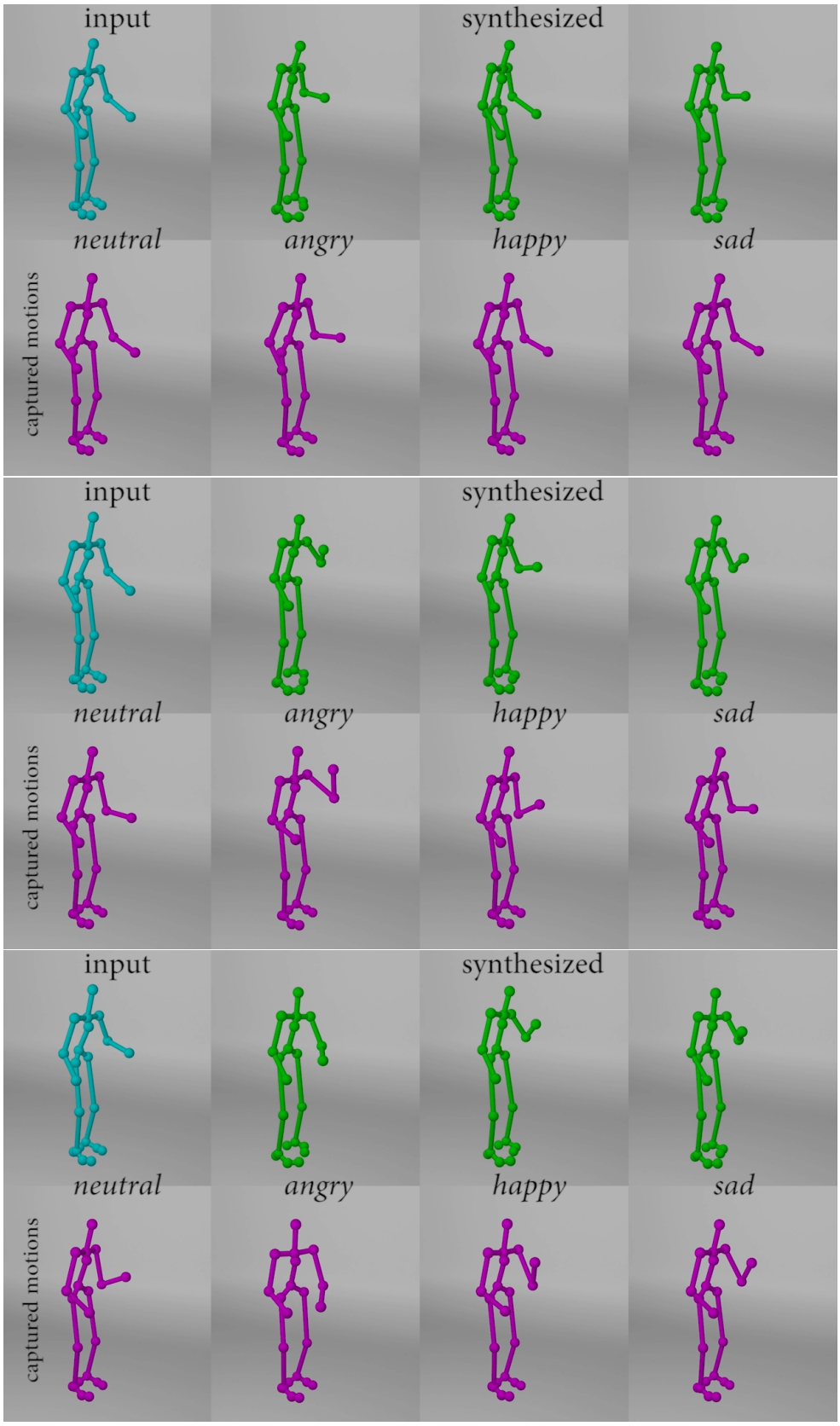


**Figure 8.** Screenshots of the *throwing* motion extracted from different stages - early (top row), middle (middle row), late (bottom row).