Emotional Facial Expression Transfer based on Temporal Restricted Boltzmann Machines

Shuojun Liu*, Dong-Yan Huang†, Weisi Lin*, Minghui Dong†, Haizhou Li† and Ee Ping Ong†
* School of Computer Engineering, Nanyang Technological University, Nanyang 639798, Singapore.
E-mail: liushj09@gmail.com, wslin@ntu.edu.sg Tel: +65-6790 6651
† Human Language Technology Department, 1 Fusionopolis way,#21-01 Connexis (South Tower), Singapore 138632
E-mail: {huang, mhdong, hli, epong}@i2r.a-star.edu.sg Tel: +65-64082639

***Abstract*—Emotional facial expression transfer involves sequence-to-sequence mappings from an neutral facial expression to another emotional facial expression, which is a well-known problem in computer graphics. In the graphics community, current considered methods are typically linear (e.g., methods based on blendshape mapping) and the dynamical aspects of the facial motion itself are not taken into account. This makes it difficult to retarget the facial articulations involved in speech. In this paper, we apply a temporal restricted Boltzmann machines based model to emotional facial expression transfer. The method can encode a complex nonlinear mapping from the motion of one neutral facial expression to another emotional facial expression which captures facial geometry and dynamics of both neutral state and emotional state.**

I. Introduction

Facial animation is to enable natural communication between human and machine. Computer facial animation can be found applications in the entertainment industry, human-computer interaction, and information retrieval systems. Therefore, various facial animation methods have been developed. In the early research stage, physically-based modeling techniques are used to generate facial expressions by simulating muscle and skin movement [1], [2], [3], [4], [5], [6], such as Facial Action Coding System (FACS) [2]. The facial emotion was decomposed into small Action Units(AUs) based on facial anatomy by FACS, which uses the movement of some specific AUs to generate different expressions.

Recently, performance-driven techniques have been widely explored [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] because the facial movement is produced based on the facial motion data of real people. Compared to physically-based modeling, performance-driven techniques can potentially generate more realistic and more natural emotion using a large face expressions database.

According to Pighin [28], performance-driven techniques can be classified into two categories. One category is called phoneme-based methods, which connect phoneme segments directly from facial motion database [11], [20], [21]. Kshirsagar and Thalmann [11] presented a method in which motion-captured data are categorized and divided into small parts based on syllable motions and new speech animation is synthesized by concatenating syllable motions from created database. In order to combine longer phoneme or syllable sequences, the search algorithms were further improved for created database [21], [22], [23]. Another category so called data-driven facial animation approaches aim to learn statistical model from motion-captured data. Chuang et al. [12], [13] learned a mapping from neutral emotion to other emotions using bilinear models. To synthesize an expressive facial animation, the facial animation with neural emotion is synthesized first and emotional animation is created through mapping. Cao *et al.* [7] proposed to learn a mapping from neutral emotion to other emotions through a Radial Basis Function (RBF) and this mapping is used to transfer neutral talk to expressive talk. Deng *et al.* [8] aligned expressive motion data with neutral motion data by phoneme-based time warping and extract pure expressive motion signals by subtracting neutral motion from expressive data. These pure expressive motion signals are used to generate expressive facial animation from neutral ones. Recently, Susskind *et al* have used restricted Bolzmann machines (RBMs) for facial expression generation, but not retargeting. They focuses on static rather than temporal data. However, none of these models explicitly incorporate dynamics into the mapping. Zeiler *et al* addressed this limitation by proposing an approach with input-output temporal restricted Boltzmann machines [29] for facial expression transfer, which can encode a complex nonlinear mapping from the motion of one individual to another.

In this paper, we apply input-output temporal restricted Boltzmann machines to emotional facial expression transfer using emotional facial motion capture data and show that the method can capture facial geometry and dynamics of both neutral state and emotional state of speaker in a natural and harmonic way. This technique can make our animated characters produce emotional facial expression and motions more efficiently.

The organization of paper is as follows. In Section II presents an emotional facial expression transfer system of which the mapping models are built from emotionally expressive motion data and neutral motion data, and can be used to generate novel synthetic emotional facial motions associated with novel neutral facial motion segments. Section III presents the recording stage and pre-processes the motion capture data. In Section IV, we briefly review several latent variable models which can map an sequence to an output sequence. We review the temporal restricted Boltzmann machine and the input-output temporal restricted Boltzmann machine (IOTRBM)

978-616-361-823-8

which extends the architecture to model an output sequence conditional on an input sequence. The simulation results will be presented against three baseline systems with IOTRBM. Finally, Section VI concludes this paper.

II. System Description for Emotional Facial Expression Transfer

Facial expression transfer, also called motion retargeting or cross-mapping, is the process of adapting the motion of the recorded (source) performance to the target character. It is related closely to very active research areas of facial motion capture and performance-driven animation over the last several years.

The main issues of this task involves facial model parametrization (called "rig" in the industry parlance) and the natural of the cross-mapping [28]. Blendshape animation is the most popular facial model parametrization technique, where a rig is a set of linearly combined facial expression each controlled by a scalar weight. The task of retargeting becomes to estimate a set of blending weights at each frame of the source data that accurately reconstructs the target frame. There are many approaches to select the blendshape from a simple selection of a set of sufficient frames from the data to blendshap models created by principal component analysis. Other common parameterizations are to simply represent the face by its vertex, polygon or spline geometry, which have many degree of freedom.

In general, the targeting problem can be formulated as a mapping function estimation problem, which produces a target expression for each source expression. A linear mapping is the simplest approximation to any mapping, and is a common choice for cross mapping. However, the inconvenience of this cross-mapping is that it cannot produce subtle nonlinear motion required for realistic graphics applications. Inspired from the IOTRBM approach for personalized facial expression transfer [29], we apply this method to our emotional facial expression transfer using marker-based motion capture data, which will drive the blendshape rig. In this paper, we focus on emotional facial expression transfer from neutral state to any emotional state using marker-based motion capture data and build models explicitly integrating dynamics into the mapping by IOTRBM.

Figure 1 illustrates our system for emotional facial expression transfer including two stages: training and testing. In training stage, the different expressive facial motion data and its accompanying speech are recorded simultaneously, forced-aligned and preprocessed. Then a temporal restricted Boltzmann machines is used to learn models from the different emotional facial motion capture data pairs such as (neutral, happy), (neutral, angry), (neutral, fear), (neutral, sad), (neutral, surprised), and (neutral, disgusted). In the testing stage, based on the learned emotional facial expression mapping functions, a new emotional facial expression sequence is generated according to a neutral emotional facial expression sequence and specified expression.

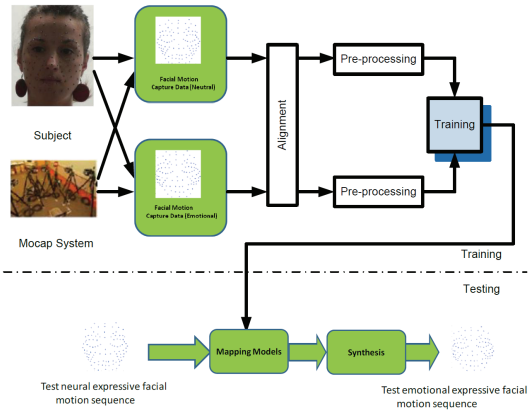

Fig. 1. Overview of emotional facial expression transfer system

III. Motion Capture Data Collection and Preprocessing

To acquire high quality data, A VICON motion capture system [35] with camera rigs was used to capture the motions of a professional actress, who had 103 markers on her face. The actress was arranged to speak an exquisitely designed corpus seven times, with different emotions, including neutral, happy, sad, angry, fear, surprised, disgusted emotions. The corpus is composed with 539 phoneme-balanced sentences. The corpus was designed to cover most of the frequently-used diphone combinations analyzed from the CMU Sphinx English dictionary. The facial motion data was captured with a rate of 100 frames per second. The actress was asked to speak the sentences with full intensity expressions. The markers' motion and aligned audio were recorded by the system simultaneously. Figure 1 illustrates the facial motion capture.

After data collection, we normalized the facial motion data. A neutral emotion frame with closed-mouth pose was chosen as a reference frame. The other data was aligned to the reference frame through translating, scaling, and rotation. We used a speech recognition engine to align recorded audio with corresponding phonemes in the forced-alignment mode. The aligned results are checked one by one to correct some errors. Then the alignment results were used to align each phoneme with its corresponding motion data segments. After that, based on phoneme, we aligned expressive captured data strictly with neutral emotion data through time warping and re-sampling.

After that, the head motion data were removed from the motion capture data. Then principle component analysis (PCA) algorithm was used to reduce the dimensionality of these motion capture vectors. We reduced the dimension to 30, covering over 99.1% of variation.

IV. Learning with Temporal Restricted Boltzmann Machines

In this section, we review a class of nonlinear generative models for high-dimensional time series. We first review a model based on the restricted Boltzmann machine (RBM) that

uses an undirected model with binary latent variables and real-valued "visible" variables. The latent and visible variables at each time step receive directed connection from visible variables at the last few time-steps. This "conditional" RBM (CRBM) makes on-line inference efficient and allows us to use a simple approximate leaning procedure. The power of the approach can be demonstrated by synthesizing various sequences from a model trained on motion capture data and by performing on-line filling in of data lost during capture. Then we review the temporal restricted Boltzmann machine and show how to generate expressive facial motion using the input-output temporal restricted Boltzmanne machine, which extends the architecture to model an output.

A. Temporal Restricted Boltzmann Machines

We have emphasized that models with distributed hidden state are necessary for efficiently modeling complex time series. But using distributed representations for hidden state in directed models of time series (Bayes nets) makes inference difficult in all but the simplest models (HMMs and linear dynamical systems). If, however, we use a restricted Boltzmann machine (RBM) to model the probability distribution of the observation vector at each time frame, the posterior over latent variables factorizes completely, making inference easy. In this section, we first review the RBM and then show a simple extension to capture temporal dependencies yet maintain its most important computational properties: simple, exact inference and efficient approximate learning using the contrastive divergence algorithm.

The restricted Boltzmann machine [32] is a Boltzmann machine with a special structure. It has a layer of visible units fully connected to a layer of hidden units but no connection within a layer. This bi-partite structure ensures that the hidden units are conditionally independent given a setting of the visible units and vice-versa. Simplicity and exactness of inference are the main advantages to using a RBM compared to a fully connected Boltzmann machine.

The RBM can be extended to model temporal data by treating its visible units and/or hidden units on a short history of inputs. Thus, this new model is called Temporal restricted Boltzmann machine [31]. Inference is more difficult than in the standard RBM if conditioning the model on the previous settings of the hidden units. Although it is possible to approximate the posterior distribution with the filtering distribution (treating the past setting of the hidden units as fixed), we still choose to use a simplified form of the model which connect only on previous visible states [34]. This model has the same advantages of the standard RBM in the computational properties; simple, exactness and efficient learning.

Typically, RBMs use stochastic binary units for both observed and latent variables. In order to model real-valued data (e.g., the parameterization of a face), we can use a modified RBM with binary logistic hidden units and real-valued Gaussian visible units [30]. This model, shown in Fig. 2 (a), defines a joint probability distribution over a real-valued representation of the current frame of data, $\mathbf{o}_t$ and a collection of binary latent variables $\mathbf{h}_t, h_j \in \{0,1\}$, where the energy function is now

$$p(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}) = \frac{\exp\left(-E(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t})\right)}{Z(\mathbf{o}_{<t})} \tag{1}$$

where $E(\mathbf{o}_t, \mathbf{h}_t)$ is an energy function. In order to simplify notation, we put a recent past of data at $t-1, \cdots, t-N$ into a vector $\mathbf{o}_{<t}$. The distribution (Eq. 1) is conditional on this recent past and normalized by a quantity Z which is a normalization constant called the partition function, whose name comes from statistical physics. The partition function is intractable to compute exactly but not needed for inference nor learning. The energy function is given by

$$E(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}) = \frac{1}{2}\sum_i (o_{i,t} - \hat{c}_{i,t})^2 - \sum_j h_{j,t}\hat{d}_{j,t} - \sum_{ij} W_{ij} o_{i,t} h_{j,t} \tag{2}$$

which models pairwise interactions between parameters and recent past, assigning high energy to improbable configurations and low energy to probable configurations. The vector $\hat{c}_{i,t}$ is a dynamically changing bias that is an affine function of the past:

$$\hat{c}_{i,t} = c_i + \sum_k C_{kj} o_{k,<t} \tag{3}$$

where k indexes the history vector. Weight matrix C and offset vector $\mathbf{c}$ (with elements c_i) parameterize the autoregressive relationship between the history and current frame of data. Each hidden unit h_j contributes a linear offset to the energy which is also a function of the history:

$$\hat{d}_{j,t} = d_j + \sum_k D_{kj} o_{k,<t} \tag{4}$$

The matrix D and the dynamical bias $\mathbf{d}$ (with elements d_i) parameterize the relationship between the history and the latent variables. The final term of Eq. 2 is a bi-linear constraint on the interaction between the current setting of the visible units and hidden units, characterized by matrix W.

The probability of observing $\mathbf{o}_t$ can be expressed by marginalizing out the binary hidden units in Eq. 1:

$$p(\mathbf{o}_t|\mathbf{o}_{<t}) = \sum_{\mathbf{h}_t} p(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}) = \sum_{\mathbf{h}_t} \frac{\exp(-E(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}))}{Z(\mathbf{o}_{<t})} \tag{5}$$

while the probability of observing a sequence, $\mathbf{o}_{N+1:T}$, given $\mathbf{o}_{1:N}$, is just the product of all the local conditional probabilities up to time T, the length of a sequence:

$$p(\mathbf{o}_{(N+1):T}|\mathbf{o}_{1:N}) = \prod_{t=N+1}^{T} p(\mathbf{o}_t|\mathbf{o}_{<t}) \tag{6}$$

The TRBM has been used to generate and denoise sequences [31], [34], as well as a prior in multi-view person tracking [33]. In all cases, the initialization is required, $\mathbf{o}_{1:N}$, to perform these tasks. Alternatively, by learning a prior model of $\mathbf{o}_{1:N}$ it could easily extended to model sequences non-conditionally, i.e., defining $\mathbf{o}_{1:T}$.

B. Input-Output Temporal Restricted Boltzmann Machines (IOTRBM)

Here we give a brief review of IOTRBM. As the objective of this paper is to transfer a neutral facial expression to an emotional facial expression by a mapping function, it is more interesting in learning a probabilistic mapping from an input sequence (, $\mathbf{s}_{1:T}$ to an output sequence, $\mathbf{o}_{1:T}$. In other words, we look for a model that defines $p(\mathbf{o}_{1:T}|\mathbf{s}_{1:T})$. However, the TRBM only defines a distribution over an output sequence $p(\mathbf{o}_{1:T})$. The idea of IOTRBM is to extend the TRBM method to learn an input-output mapping [29]. In addition to having access to the complete history of the input, the first N frames of the output are accessed to seek a model $p(\mathbf{o}_{(N+1):T}|\mathbf{o}_{1:N}, \mathbf{s}_{1:T})$. In steady of an Nth order Markov assumption on the current output, $\mathbf{o}_t$, that is, assuming conditional independence on all other variables given an N-frame history of $\mathbf{o}_t$ and an $N+1$-frame history of the input (up to and including time, t), the model $p(\mathbf{o}_{(N+1):T}|\mathbf{o}_{1:N}, \mathbf{s}_{1:T})$ can be expressed in an online setting:

$$p(\mathbf{o}_{(N+1):T}|\mathbf{o}_{1:N}, \mathbf{s}_{1:T}) = \prod_{i=N+1}^{T} p(\mathbf{o}_t|\mathbf{o}_{<t}, \mathbf{s}_{<=t}) \quad (7)$$

where the shorthand $\mathbf{s}_{<=t}$ is used to describe a vector that concatenates a window over the input at time $t, t-1, \cdots, t-N$.

The TRBM can be easily adapted to model $p(\mathbf{o}_{(N+1):T}|\mathbf{o}_{1:N}, \mathbf{s}_{1:T})$ by modifying its energy function to incorporate the input. The general form of energy function is still the same as Eq. 2 with conditioned on $\mathbf{s}_{<=t}$. The dynamic biases (Eq. 3 and 4) are redefined as follows:

$$\hat{c}_{i,t} = c_i + \sum_k C_{kj}\mathbf{o}_{k,<t} + \sum_l P_{li}\mathbf{s}_{l,<=t} \quad (8)$$

$$\hat{d}_{j,t} = d_j + \sum_k D_{kj}\mathbf{o}_{k,<t} + \sum_l Q_{li}\mathbf{s}_{l,<=t} \quad (9)$$

where l is an index order over elements of the input vector. Therefore the matrix P ties the input linearly to the output (much like existing simple models) but the matrix Q also allows the input to nonlinearly interact with the output through the latent variables $\mathbf{h}$. This model is called an Input-Output Temporal Restricted Boltzmann Machine (IOTRBP) [29]. It is depicted in Fig. 2 (b).

A desirable criterion for training the model is to maximize the conditional log likelihood of the data:

$$\mathcal{L} = \sum_{t=N+1}^{T} \log p(\mathbf{o}_t|\mathbf{o}_{<t}, \mathbf{s}_{<=t}) \quad (10)$$

It is obviously that the gradient of Eq. 10 with respect to the model parameters $\theta = \{W, C, D, P, Q, \mathbf{c}, \mathbf{d}\}$ is difficult to compute analytically due to the normalization constant Z. Contrastive Divergence (CD) learning is typically used in steady of maximum likelihood. It follows the approximate gradient of an objective function that is the different between

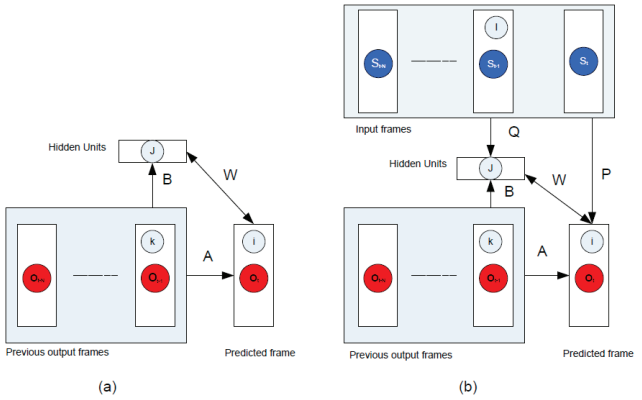

Fig. 2. (a) A Temporal Restricted Boltzmann Machine. (b) An Input-Output Temporal Restricted Boltzmann Machine.

two Kullback-Leibler divergence [36]. It is widely used in practice and tends to produce good generative models [37].

The CD updates for the IOTRBM have a common form (see the supplementary material for details):

$$\begin{aligned} \delta\theta_i \quad \propto \quad & \sum_{t=N+1}^{T} < \frac{\partial E(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}, \mathbf{s}_{<=t})}{\partial \theta_i} >_{\text{data}} \\ - \quad & < \frac{\partial E(\mathbf{o}_t, \mathbf{h}_t|\mathbf{o}_{<t}, \mathbf{s}_{<=t})}{\partial \theta_i} >_{\text{recon}} \end{aligned} \quad (11)$$

where $<>_{\text{data}}$ is an expectation with respect to the training data distribution, and $<>_{\text{recon}}$ is the M-step reconstruction distribution as obtained by alternating Gibbs sampling, starting with the visible units clamped to the training data. The input and output history stay fixed during Gibbs sampling. In CD, two main operations are required. Firstly, the latent variables need to be sampled, given a window of the input and output,

$$p(h_{i,t} = 1|\mathbf{o}, \mathbf{o}_{<t}, \mathbf{s}_{<=t}) = \frac{1}{1 + \exp(-d_{j,t} - \sum_i W_{ij}\mathbf{o}_{i,t})} \quad (12)$$

Secondly, the output data are reconstructed given the latent variable

$$p(o_{i,t}|\mathbf{h}_t, \mathbf{o}_{<t}, \mathbf{s}_{<=t}) = \mathcal{N}(o_{i,t}; \hat{c}_{i,t} + \sum_j W_{ij}h_{j,t}, 1) \quad (13)$$

Eq. 12 and 13 are alternated M times to arrive at the M-step quantities used in the weight updates.

V. Experiments

The purpose of this paper is to study the performance of the IOTRM on emotional facial expression transfer.

We evaluate the IOTRBM on emotional facial expression transfer dataset: 2D motion capture data of neutral and happy facial expressions. We compare IOTRBM with the three methods such as linear regression (LR), Nth-order Autoregressive model (AR), and multilayer perception (MLP) on the dataset.

For aligned neutral-happy pair of sequences, a regularized linear regression (LR) is performed between each frame of the input (neutral sequence) to each frame of the output (emotional sequence). The least squares method is applied to analytically

solve the model. The regularization parameter is set by cross-validation on the training set.

Nth-order Autoregressive model (AR) is used to improve the linear regression model by integrating linear dynamics through the history of the input and output. The regularized least squares method is used to fit a matrix that maps from a concatenation of the $(N+1)$-frame input window $s_{<=t}$ and N-frame target window, $o_{<t}$.

Multilayer perception (MLP) is a nonlinear model with one deterministic hidden layer, the same cardinality as the IOTRBM. The concatenation of the source and target history is taken as the input, the current target frame is considered as the output. A nonlinear conjugate gradient method is used to train the datasets.

These methods were chosen to compare to show the main differences of IOTRBM approach over the majority of techniques proposed for this application, namely the consideration of dynamics and the use of a nonlinear mapping through latent variables.

During the experiments, we set all models with a window of 4 input frames (3 previouse + 1 current) and 6 previous output frames as in [29], with the exception of linear regression which only saw the current input. For the IORBM models, if the parameters C and P are initialized to the solution found by the autoregressive model, slightly better results can be obtained. All other parameters were initialized to small random values. For constractive divergence (CD) learning we set the learning rates for C and P to 10^{-6} and for all other parameters to 10^{-3} in order to prevent strong correlations from dominating early in learning. All parameters used a fixed weight decay of 0.025 and momentum of 0.70. A small amount of Gaussian noise ($\sigma = 0.1$) is added to the output history to make the model more robust to unseen outputs during prediction based on the suggestion by [33].

A. Alignment and pre-processing

We consider a dataset which consists of emotional facial motion capture data of an actress who was asked to speak the same sentences in different emotional states It has 532 trials for different emotions. Each frame is 206 dimensional, representing the x and y positions of facial markers on the face. We selected 200 trials for different emotions.

The FESTIVAL system [43] was used to perform phoneme-alignment by aligning each phoneme with its corresponding motion capture segments. This alignment work was done by inputting audio and its accompanying text scripts into the speech recognition program in a force-alignment mode. We form different emotional pairs such like (neutral, happy), (neutral, angry), (neutral, sad), (neutral, disgusted), and (neutral, surprised).

Each pair of sequences has been manually time-aligned based on a phonetic transcription so they are synchronized between emotion states.

We found the original data to exhibit significant random relative motion between the two emotional faces of a subject through the entire sequences which could not reasonable be modeled. Therefore, we transformed the data with an affine transform on all markers in each frame such that a select few nose and skull markers per frame (stationary facial locations) were approximately fixed relative to the first frame of the neutral sequences. For each pair of sequences, both the input and output were reduced to 30 dimensions by retaining only their first 30 principal components. This maintained 99.9% of the variance In the data. Finally, the data was normalized to have zero mean and scaled by the average standard deviation of all the elements in the training set.

We evaluate the various methods on 6 random arbitrary splits of the dataset of emotional pair (neutral, happy). In this case, 150 complete sequences are maintained for training and the remaining 50 sequences are used for testing. Each model is presented with the first 6 frames of the true test output and successive 4-frame windows of the true test input. The exception is the linear regression model, which only sees the current input. Therefore prediction is measured from the 7th frame onward.

B. IOTRBM setting

The IOTRBM produces its final output by initializing its visible units with the current previous frame plus a small amount of Gaussian noise and then performing 30 alternating Gibbs steps. At the last step, the hidden units are not sampled. This predicted output frame now becomes the most recent frame in the output history and we iterate forward. The results show a IOTRBM with 3 hidden units. A model with different hidden (10, 50, 100) units are tried and showed to perform slightly worse.

We report RMS marker error in mm where the mean is taken over all markers, frames and test sequences (Table V-B) for the sequences pairs (neutral, happy). We can observe that the IOTRBM consistently outperforms linear regression. In all but two splits (where performance is comparable) the IOTRBM outperforms the AR model. Mean performance over the splits show an advantage to IOTRBM.

The robustness of each model are compared by corrupting inputs or outputs. Various amounts of white Gaussian noise are added to the input window, output history initialization or both during retargeting with a trained model. The performance of each model is given in Table V-B. The IOTRBM generally outperforms the baseline models in the presence of noise. This is most apparent in the case of input noise: the scenario we would most likely find in practice. However, under low to moderate output noise, it is worthy to note that the IOTRBM is robust for any N frame output initialization to produce a sensible retargeting.

VI. Conclusions

This paper applied a type of temporal restricted Boltzmann machines (TRBM) to learn to generate happy facial expression from neutral expression for an intelligent dialogue expressive avatar. The TRBM is an extension of the RBM to model temporal data by conditioning its hidden units only on previous visible states. This simplified TRBM inherits the most

TABLE I
2D DATASET, MEAN RMS ERROR ON TEST OUTPUT SEQUENCES (NEUTRAL, HAPPY)

	RMS Marker Error (mm)						
Model	S1	S2	S3	S4	S5	S6	mean
Linear regression	7.04	7.02	7.03	7.24	6.96	6.64	6.99 $\pm$ 0.14
Autoregressive	6.28	**6.07**	6.51	6.14	6.13	6.41	6.25 $\pm$ 0.17
MLP	6.12	6.09	6.50	6.02	6.05	6.03	6.06 $\pm$ 0.09
IOTRBM	**6.10**	6.08	**5.80**	**5.72**	**6.02**	**5.98**	**5.95 $\pm$ 0.11**

TABLE II
2D DATASET, MEAN RMS ERROR (IN MM) UNDER NOISY INPUT AND OUTPUT HISTORY (SPLIT 4)

	Input noise			Output noise			Input & Output noise		
Model	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
Linear regression	7.04	16.02	147.03	N/A					
Autoregressive	6.85	11.23	89.41	7.64	8.13	39.51	6.25	12.03	97.23
MLP	6.50	6.69	7.20	6.23	6.45	6.32	6.26	6.45	8.03
IOTRBM	**6.23**	**6.18**	**6.22**	**6.02**	**6.00**	9.43	**6.22**	**6.32**	8.50

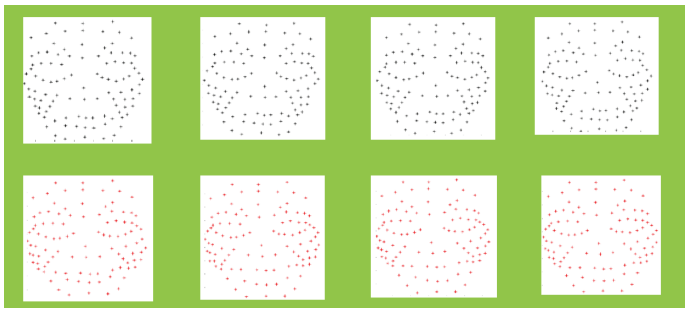

Fig. 3. retargeting with IOTRBM. The top row are the input (blue stars). The bottom row shows the prediction from our model (red stars

important computational properties of RBMs: simple, exact inference and efficient approximate learning of non-linear structure and dynamics. The experimental results show that the TRBM is powerful enough not only to capture the large range of variations in expressive appearance in static face, but also to capture the dynamics in temporal data. We shall apply this method to other 2D and 3D emotional pairs and improve further the performance of facial expression transfer as all the existing facial expression transfer are unable to provide human-like facial expression in the retarget motion. We are interested in exploring the extensions of the model including style-based contextual variables as well as integrating the technique into our intelligent dialogue expressive avatar for facial animation.

REFERENCES

[1] K. Kahler, J. Haber, and H. P. Seidel, "Geometrybased muscle modeling for facial animation," In *Proceedings of Graphics Interface,* pp. 37 - 46, 2001.

[2] P. Ekman and W. V. Friesen, Facial action coding system: a technique for the measurement of facial movement, *Consulting Psychologists Press,* Palo Alto, CA., 1978.

[3] P. Kalra, A. Mangili, N. M. Thalmann, and D. Thalmann, Simulation of facial muscle actions based on rational free from deformations, *Eurographics,* vol. 11, no. 3, pp. 59 - 69, 1992.

[4] K. Kahler, J. Haber, and H. P. Seidel, Physically-based facial modeling, analysis, and animation, *Proc. Graphics Interface Conf.,* 2001.

[5] D. Terzopoulos and K. Waters, Physically-based facial modeling, analysis, and animation, *J. Visualization and Computer Animation,* vol. 1, no. 4, pp. 73 - 80, 1990.

[6] Y. Tang, M. Xu, and Z. Cai, Research on facial expression animation based on 2d mesh morphing driven by pseudo muscle model, *International Conference on Educational and Information Technology (ICEIT),* 2010, vol. 2, pp. 403 - 407, 2010.

[7] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, Expressive speechdriven facial animation, *ACM Transactions on Graphics,* vol. 24, no. 4, pp. 1283 - 1302, Oct 2005.

[8] Z. Deng, U. Neumann, J. P. Lewis, T. Y. Kim, M. Bulut, and S. Narayanan, Expressive facial animation synthesis by learning speech co-articulations and expression spaces, *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, no. 6, pp. 1523 - 1534, 2006.

[9] B. Brand, Voice puppetry, In *Proceedings of ACM SIGGRAPH Conference,* pp. 21 - 28, 1999.

[10] T. Ezzat, G. Geiger, and T. Poggio, Trainable videorealistic speech animation,, *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 388 - 398, 2002.

[11] S. Kshirsagar and N. M. Thalmann, Visyllable based speech animation, *Computer Graphics Forum (Proc. Eurographics Conf.),* vol. 22, no. 3, 2003.

[12] E. Chuang and C. Bregler, Moodswings: Expressive speech animation, *ACM Transactions on Graphics,* vol. 24, no. 2, pp. 331 - 347, 2005.

[13] E. Chuang, H. Deshpande, and C. Bregler, Facial expression space learning, In *Proceedings of Pacific Graphics 2002*, pp. 68 - 76, 2002.

[14] Z. Deng and U. Neumann, Expressive speech animation synthesis with phoneme-level controls, *Computer Graphics Forum,* vol. 27, no. 8, pp. 2096 - 2113, 2008.

[15] Z. Deng, J. P. Lewis, and U. Neumann, Automated eye motion synthesis using texture synthesis, *IEEE Computer Graphics and Applications,* vol. 25, no. 2, pp. 24 - 30, 2005.

[16] G. Kalberer and L. V. Gool, Face animation based on observed 3d speech dynamics, *Proc. IEEE Computer Animation Conf.,* pp. 20 - 27, 2001.

[17] P. Cosi, C. E. Magno, G. Perlin, and C. Zmarich, Labial coarticulation modeling for realistic facial animation, *Proc. Intl Conf. Multimodal Interfaces,* pp. 505 - 510, 2002.

[18] S. A. King and R. E. Parent, Creating speech-synchronized animation, *IEEE Trans. Visualization and Computer Graphics,* vol. 11, no. 2, pp. 341 - 352, 2005.

[19] C. S. Chan and F. S. Tsai, Computer Animation of Facial Emotions, in *International Conference on Cyberworlds (CW),* 2010, pp. 425 - 429.

[20] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, Accurate automatic visible speech synthesis of arbitrary 3d model based on concatenation of diviseme motion capture data, *Computer Animation and Virtual Worlds,* vol. 15, pp. 1 - 17, 2004.

[21] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, Real-time speech motion synthesis from recorded motions, In *Proceedings of Symposium on Computer Animation,* pp. 345 - 353, 2004.

[22] E. Cosatto and H. P. Graf, Audio-visual unit selection for the synthesis of photo-realistic talking-heads, In *Proceedings of ICME,* pp. 619 - 622, 2000.

[23] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, Accurate automatic visible speech synthesis of arbitrary 3d model based on concatenation of diviseme motion capture data, *IEEE Transactions on Visualization and Computer Graphics,* vol. 15, no. 5, pp. 485 - 500, 2005.

[24] A. Niswar, E. P. Ong, H. T. Nguyen, and Z. Huang, Real-time 3D Talking Head from a Synthetic Viseme Dataset, in *Proceedings of Virtual-Reality Continuum and its Applications in Industry (VRCAI),* ACM, pp. 29 - 33, 2009.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing,* Cambridge University Press, 1992.

[26] L. Liang, C. Liu, Y Q. Xu, B. Guo, and H. Y. Shum, Real-time texture synthesis by patch-based sampling, *ACM Transactions on Graphics,* vol. 20, no. 3, pp. 127 - 150, 2001.

[27] H. T. Nguyen, E. P. Ong, A. Niswar, Z. Huang, and S. Rahardja, Automatic and real-time 3d face synthesis, in *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry* 2009, ACM, 2009, pp. 103 - 106.

[28] F. Pighin and J. P. Lewis, "Facial motion retargeting," In *ACM SIGGRAPH 2006 Courses, SIGGRAPH* 06, New York, NY, USA, 2006. ACM.

[29] M. D. Zeiler, G. W. Taylor, L. Sigal, I. Matthews, and R. Fergus, "Facial Expression Transfer with Input-Output Temporal Restricted Boltzmann Machines," *NIPS,* 2011, pp.1629 - 1637.

[30] Y. Freund and D. Haussler, "Unsupervised learning of distributions of binary vectors using 2-layer networks," In *Proc. NIPS 4,* 1992.

[31] I. Sutskever and G. Hinton, "Learning multilevel distributed representations for high dimensional sequences," In *Proc. AISTATS,* 2007.

[32] P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, et al., editors, Parallel Distributed Processing: Volume 1: Foundations, pages 194 - 281. MIT Press, Cambridge, MA, 1986.

[33] G. W. Taylor, and G. E. Hinton, "Factored conditional restricted Boltzmann machines for modeling motion style," In *Proceedings of the 26th Annual International Conference on Machine Learning,* pp. 1025 - 1032, 2009.

[34] G. W. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," In *Proc. NIPS 19,* 2007.

[35] B. Bodenheimer, C. F. Rose, S. Rosenthal, and J. Pella, "The process of motion capture: Dealing with the data," in *Pro. of Eurographics Workshop on Computer Animation and Simulation,* 1997.

[36] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation,* vol 14:1771 - 1800, 2002.

[37] M. Carreira-Perpinan and G. Hinton, "On contrastive divergence learning," In *AISTATS,* pp. 59 - 66, 2005.

[38] A. Krizhevsky, S. Ilya, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems,* vol 25, pp. 1097 - 1105, 2012.

[39] G.E. Dahl, T.N. Sainath, and G.E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," In *IEEE 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 8609 - 8613, 2013.

[40] J. Dean, G. Corrado, R. Monga, C. Kai, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q.V. Le, and A.Y. Ng, "Large scale distributed deep networks," In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q.Weinberger, editors, *Advances in Neural Information Processing Systems,* vol. 25, pp. 1223 - 1231, 2012.

[41] N. Srivastava, "Improving Neural Networks with Dropout," Master's thesis, Univeristy of Toronto, Toronto, Canada, January 2013.

[42] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," in *CoRR, abs/1207.0580,* 2012.

[43] FESTIVAL, "http://www.cstr.ed.ac.uk/projects/festival/."