Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach

Chung-Cheng Chiu[1], Louis-Philippe Morency[2], and Stacy Marsella[3]

[1] Google Inc. redjava@gmail.com
[2] Language Technology Institute, School of Computer Science, Carnegie Mellon University
morency@cs.cmu.edu
[3] Northeastern University stacymarsella@gmail.com

Abstract. Gestures during spoken dialog play a central role in human communication. As a consequence, models of gesture generation are a key challenge in research on virtual humans, embodied agents capable of face-to-face interaction with people. Machine learning approaches to gesture generation must take into account the conceptual content in utterances, physical properties of speech signals and the physical properties of the gestures themselves. To address this challenge, we proposed a gestural sign scheme to facilitate supervised learning and presented the DCNF model, a model to jointly learn deep neural networks and second order linear chain temporal contingency. The approach we took realizes both the mapping relation between speech and gestures while taking account temporal relations among gestures. Our experiments on human co-verbal dataset shows significant improvement over previous work on gesture prediction. A generalization experiment performed on handwriting recognition also shows that DCNFs outperform the state-of-the-art approaches.

1 Introduction

Embodied conversational agents (ECAs) are virtual characters capable of engaging face-to-face interaction with human and play an important role in many applications such as human-computer interaction [6] and social skills training [29]. A key challenge in building an ECA is giving them the ability to use appropriate gestures while speaking, as users are sensitive to whether the gestures of an ECA are consistent with its speech [11]. This challenge is also true for social robotic platforms [30]. Such co-verbal gestures [36] must coordinate closely with the prosody and verbal content of the spoken utterance. Manual development of an agent's gestures is typically a tedious process of manually handcrafting gestures and assigning them to the agent's utterances. A data-driven approach that learns to predict and generate co-verbal gestures is a promising alternative to such manual approaches.

However, the prediction and generation of co-verbal gestures presents a difficult, novel machine learning challenge in that it must span and couple multiple domains: the conceptual content in the utterance, utterance prosody and the physical domain of gestural motions. The coupling between these domains has several complex features. There is a tight coupling between gesture motion, the evolving the content of the utterance as well as the prosody of speech. This coupling is the product of the information

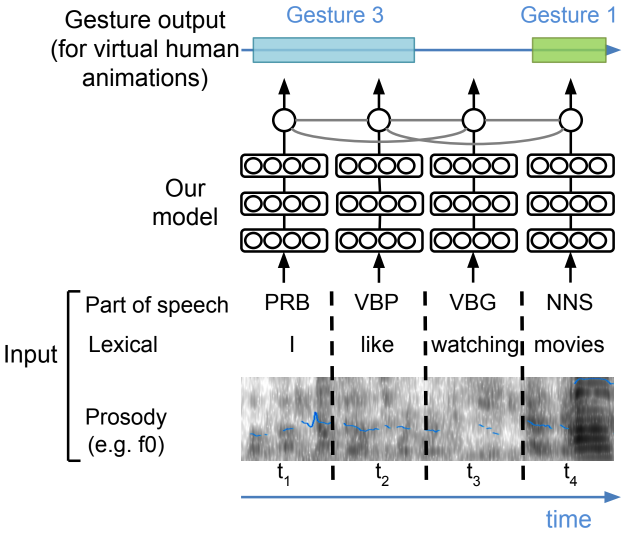

Fig. 1: The overview of our framework for predicting co-verbal gestures. Our Deep Conditional Neural Field (DCNF) model predicts gestures by integrating verbal and acoustic while preserving the temporal consistency.

conveyed through both speech and gestures [4] that may be shared at a hidden, abstract level [25] which relates utterance content and physical gestures. These properties suggest that generating gestures from speech can exploit a representation that takes into account this relation between form and function (what the gesture conveys) and a model capable of modeling the deep and temporal relationship between speech and gestures. Additionally, speech and gesture are closely coupled in time, which raises its own challenges since gestures are physical motions with tight temporal and spatial constraints if the motion is to look natural.

In this paper, we introduce a deep, temporal model to realize the prediction of gestures from verbal content and prosody of the spoken utterance. The structure of the entire framework is shown in Figure 1. Our model, called deep conditional neural field (DCNF), is an extension of previous work [10, 13] that combines the advantages of deep neural network for mapping complex relation and an undirected second-order linear-chain for modeling the temporal coordination of speech and gestures. We also propose a gesture representation scheme that takes advantage of previous literature that relates the form and communicative function of gestures [18, 4, 24].

We assess our framework by evaluating the prediction accuracy on actual co-verbal gesture prediction data involving dyadic interviews, showing that our model outperforms state-of-the-art approaches.

2 Related Work

Data-driven approaches to generate co-verbal gestures for intelligent embodied agent have received increasing attention in gesture research. [32] took the co-generation perspective in which the framework synthesizes both speech and gestures based on the determined utterance during the conversation. [27] addressed modeling individual gesture styles through analyzing the relation in the data between extracted utterance information and a person's gestures. Our technique can be applied to predict this information, and their approaches can then be applied to accomplish the gesture generation process. [19] also took the co-generation perspective and focused on modeling individual styles on iconic gestures to improve human-agent communication.

Some of the previous work focused on realizing the relation between prosody and motion dynamics [23, 22, 8]. By using only prosody as input, these models do not require speech content analyses but are limited to the subset of gestures that correlate closely to prosody, for example, a form of rhythmic gesture called beats. Our approach goes beyond prosody to realize a mapping from the utterance content to more expressive gestures and can be integrated to extend existing work to generate animations beyond beat gestures.

Alternatives to data-driven machine learning approaches are the handcrafted rule-based approaches [21, 7, 24, 1]. These exploit expert knowledge on speech and gestures to specify the mapping from utterance features to gestures. While earlier works based on this approach have focused on addressing the mapping relation between only linguistic features and gestures [21, 7], recent work [24] has also addressed how to use acoustic features to help gesture determination.

Realizing a mapping from speech to gestures involves learning a model that relates two sequences, the speech input sequence and the gesture output sequence. Recent advances in neural networks toward modeling the two sequence problems apply recurrent neural networks (RNNs) [33] and its extension, long short-term memory (LSTM) network [16]. The RNN-based architecture is designed to address problems in which the input and output time series can have different lengths and are correlated as whole sequences but may not have a strong correlation at the frame-by-frame level. The resulting model utilizes less of the structure in the data and make predictions by maximizing only the distribution of targeting sequences. On the other hand, our approach utilizes the fine-grained synchronization between observed and predicting sequences and also learns the global conditional distributions of both sequences to further improve the prediction accuracy.

Previous approaches in deep learning that utilize the synchronized structure of two sequences trained separately a deep neural network and a linear-chain graphical model. For example, in speech recognition [26] the common approach is to train deep learning with individual frames and then applies hidden Markov models (HMMs) with the hidden states. Our approach learns both the deep neural network and temporal contiguity of CRFs with a joint likelihood. There are previous works that adopt similar perspective on extending CRFs with deep structure [38, 10] and show improvement over a single-layer CRFs or CRFs combined with a shallow layer of neural network [28]. Our experiments show improvement over these approaches.

To our knowledge, this work is the first to introduce a gesture representation scheme that relates the form and communicative function of gestures and a deep, temporal model capable of realizing the relation between speech and the proposed gesture representation. [8] adopt the concept of unsupervised training of deep belief net [35], but without an effective gesture representation and a supervised training phase the learning task is much more challenging and therefore has been limited to realizing the relation between prosody and rhythmic movement. Our proposed model goes beyond prior work [10, 13] by combining the advantages of deep neural network for mapping complex relation with an undirected second-order linear-chain for modeling the temporal coordination of speech and gestures.

3 Predicting Co-Verbal Gestures

Predicting co-verbal gestures brings together many core domains of artificial intelligence, including the conceptual content in the utterance, utterance prosody and the physical domain of gestural motions. A common function of the parallel use of speech and gesture is to convey meaning in which gesture plays the complementary or supplementary role [14], and gestures may help to convey complex representations through expressing complementary information about abstract concepts [25]. Realizing this relation between speech and gesture requires realizing the hidden abstract concept. To build a successful predictive model it is important to first create a formal representation of its output label, the co-verbal gestures. Based on this idea, we exploit gestural signs [4] which summarize the functions and forms of co-verbal gestures to allow the predictions of gestures from speech signals, including utterance content and prosody. In particular, we focus on gesture categories that can be more reliably predicted from the utterance content and prosody: abstract deictic, metaphoric, and beat gestures. Abstract deictic gestures are pointing movements that indicate an object, a location, or abstract things which are not physically present in the current surroundings. Metaphoric gestures exhibit abstract concept as having physical properties. Beat gestures are rhythmic actions synchronized with speech and they tend to correlate more with prosody as opposed to utterance content. This ignores those gestures that convey information that is uncoupled or distinct from the utterance content and prosody [5] in the sense that learning would require additional information to predict the gestural signs.

We design our dictionary of gestural signs based on previous literature in gestures [18, 4, 24] and the three gesture categories, and then calculated their occurrences in a motion capture data [12] which records co-verbal gestures performed during face-to-face conversations to filter out those that rarely appeared. The final set of gestural signs has size of 14, and the list and their descriptions are shown in Table 1. This discrete set of co-verbal gestures was selected to include considerable coverage while keeping a clear distinction between gesture labels to make learning feasible. An important challenge for predicting gestural signs is to model the temporal coordination between speech and gestural signs. A state-of-the-art work [22] applies conventional conditional random fields (CRFs) for learning co-verbal gesture predictions. The limitation of conventional CRFs is that it requires defining functions for modeling the correlation between input signals and labels, and manually defining these functions that may express the relation between high-dimensional speech signals and gestures is no trivial task. Thus, we argue instead to use a deep model to learn this complex relation.

Gestural signs	Description
Rest	Resting position of both hands.
Palm face up	Lift hands, rotate palms facing up or a little bit inward, and hold for a while.
Head nod	Head nod without arm gestures.
Wipe	Hands start near (above) each other and move apart in a straight motion.
Whole	Move both hands along outward arcs with palms facing forward.
Frame	Both hands are held some inches apart, palms facing each other, as if something is between hands.
Dismiss	Hand throws to the side in an arc as if chasing away.
Block	Hand is positioned in front of the speaker, palm toward front.
Shrug	Hands are opened in an outward arc, ending in a palm-up position, usually accompanied by a slight shrug.
More-Or-Less	The open hand, palm down, swivels around the wrist.
Process	Hand moves in circles.
Deictic.Other	Hand is pointing toward a direction other than self.
Deictic.Self	Points to him/herself.
Beats	Beats.

Table 1: A formalized representation of co-verbal gestures for computational prediction.

4 Deep Conditional Neural Fields

In this section, we formally describe the Deep Conditional Neural Field (DCNF) model which combines state-of-the-art deep learning techniques with the temporal modeling capabilities of CRFs for predicting gestures from utterance content and prosody (see Figure 2). The prediction task takes the transcript of the utterance, part-of-speech tags of the transcript, and prosody features of the speech audio as input $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, and learn to predict a sequence of gestural signs $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ in which the sequence has length N. At each time step t, the gestural sign y_t is contained in the set of our gestural sign dictionary $y_t \in \mathbf{Y}$ defined in the previous section (see Table 1) , and the input x_t is a feature vector $x_t \in \mathbf{R}^d$ where d corresponds to the number of input features (see next section for a detailed description of our input features).

Following the formalism of [10] and [13], the DCNF extends previous models to follow a 2^{nd}-order Markov assumption and is defined as:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) =& \frac{1}{Z(\mathbf{x})} \sum_{t=1}^{N} \exp[\sum_{k} \theta_k^{g_1} g_k^1(y_{t-1}, y_t) \\ &+ \sum_{l} \boldsymbol{\theta_l}^{\mathbf{g_2}} \mathbf{g_l^2}(\mathbf{y_{t-1}}, \mathbf{y_t}, \mathbf{y_{t+1}}) \\ &+ \sum_{i} \theta_{i,y_t}^{f} f_i(x_t, \theta^w)] \end{aligned}$$

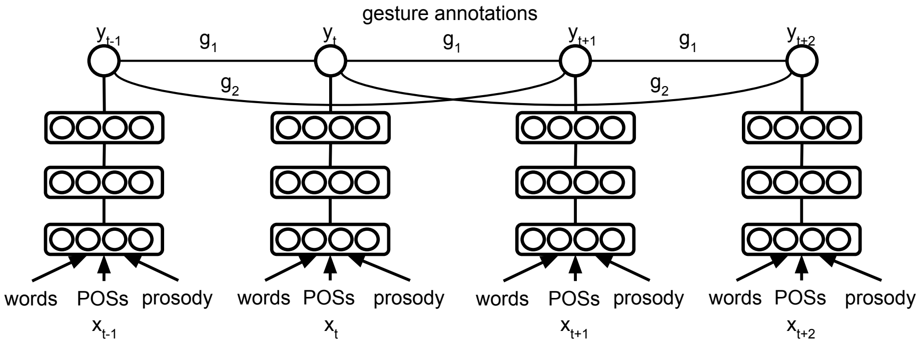

Fig. 2: The structure of our DCNF framework. The neural network learns the nonlinear relation between speech features and gestural signs. The top layer is a second-order undirected linear-chain which takes the output of the neural network as input and model the temporal relation among gestural signs. Both the top undirected chain and deep neural networks are trained jointly.

where model parameters $\boldsymbol{\theta} = [\theta^{g_1}, \theta^{g_2}, \theta^{f}, \theta^{w}]$ and $Z(\mathbf{x})$ is the normalization term. gs correspond to edge features in which $g^1(y_{t-1}, y_t)$ and $g^2(y_{t-1}, y_t, y_{t+1})$ denote the first and second order edge functions, and θ^{g_1} and θ^{g_2} correspond to their parameters respectively. The 2nd-order term $g^2(y_{t-1}, y_t, y_{t+1})$ is one of the the major improvement of the DCNF model. f is related to neural networks in which $f(x_t, \theta^w)$ associates the output of the last layer of the deep neural network with θ^f denotes its parameters, and y and $\theta^w = \{\theta^w_1, \theta^w_2, \ldots, \theta^w_{m-1}\}$ represents the network connection parameters of the m neural network layers:

$$\begin{aligned} f(x_t, \theta^w) =& h(a_{m-1}\theta^w_{m-1}) \text{ where} \\ a_i =& h(a_{i-1}\theta^w_{i-1}),\ i = 2 \ldots m-1 \end{aligned}$$

where a_i represents the output at ith neural network layer, θ^w_i represents the connection weights between ith and $i+1$th layers, and h is the activation function. This work applies the logistic function $(1/1 + \exp(-a\theta^w))$ as the activation function[4]. Readers can refer to [10, 13] for more background about the combination of CRFs and neural networks.

[4] We have experimented with both the logistic and the rectified linear $(\max(a\theta^w, 0))$ functions with similar results. Because of space constraints, we are focusing on the logistic function.

Prediction Given a sequence $\mathbf{x}$ and parameters learned from the training data, the prediction process of DCNFs predicts the most probable sequence $\mathbf{y}*$:

$$\begin{aligned}\mathbf{y}* &= \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x};\theta^{g_1},\theta^{g_2},\theta^{f},\theta^{w})\\ &= \arg\max_{\mathbf{y}} \frac{1}{Z(\mathbf{x})}\sum_{t=1}^{N}\exp[\sum_k \theta_k^{g_1} g_k(y_{t-1},y_t)\\ &+ \sum_l \theta_l^{g_2} g_l(y_{t-1},y_t,y_{t+1})\\ &+ \sum_i \theta_{i,y_t}^{f} f_i(x_t,\theta^{w})]\end{aligned}$$

To estimate the probability of each label of frame t, the neural networks take the input x_t and forward the value through the network to generate f_i, the undirected linear chain performs forward-backward belief propagation to calculate the values of g_k and g_l, and the potential of each label is the weighted summation of g_1, g_2, f and the probability of each label is its normalized potential.

Learning To prevent the overfitting of DCNFs, the model has a regularization term for all parameters and we define our objective function as follows:

$$L(\boldsymbol{\theta}) = \sum_{t=1}^{N} \log P(y_t|x_t;\boldsymbol{\theta}) - \frac{1}{2\gamma^2}\|\boldsymbol{\theta}\|^2,$$

in which $\boldsymbol{\theta}$ denotes the set of model parameters and γ corresponds to regularization coefficients. The regularization term on training the deep neural networks encourages the weight decay which reduce the complexity increase of the network connections along the parameter updates. We applied stochastic gradient descent for training DCNFs with a degrading learning rate to encourage the convergence of the parameter updates[5].

To also help prevent co-adaptation of network parameters which result overfitting, we apply the dropout technique [17] to change the feed-forward results of $f_i(x_t,\theta^w)$ in the training phase. By performing dropout, at the feed-forward phase the output of each hidden node has a probability of being disabled. Consequently the output of hidden nodes in the training phase is different from that of the testing phase. The dropout nodes are re-sampled at every feed-forward process. This stochastic behaviors encourage hidden nodes to model distinct patterns and therefore further prevent the overfitting. The dropout technique is not applied during the testing phase.

Gradient calculation To learn our model parameters, we derived the gradient of our objective function with respect to $\theta^{g_1},\theta^{g_2},\theta^{f},\theta^{w}$. We derive $\theta^{g_1},\theta^{g_2},\theta^{f}$ following previous work on CRFs [20], and derive θ^{w} with backpropagation [10, 13]. Backpropagation decomposes the gradient at each layer as the product of an error term δ with

[5] The full derivation of the gradient was omitted because of space constraint.

the input and propagates δ to the lower layers to facilitate gradient calculation. Thus, performing backpropagation on DCNF requires determining δ_{m-1} of θ^w_{m-1} in which $\nabla\theta^w_{m-1} = \delta_{m-1}\hat{a}_{m-1}$ for $\nabla\theta^w_{m-1}$ denotes the gradient of θ^w_{m-1} and $\hat{a}_{m-1}$ denotes the output at layer $m-1$ with dropout. As the gradient of θ^w_{m-1} is given by:

$$\begin{aligned}
&\frac{\partial \log P}{\partial \theta^w_{m-1}} \\
&= \sum_t^N \sum_i [\lambda_{i,y_t} \frac{\partial f_i(x_t, \theta^w)}{\partial \theta^w_{m-1}} - \sum_{\tilde{y}} p(\tilde{y}|x_t)\lambda_{i,\tilde{y}} \frac{\partial f_i(x_t, \theta^w)}{\partial \theta^w_{m-1}}] \\
&= \sum_t^N \sum_i [\lambda_{i,y_t} \frac{\partial h(\hat{a}_{m-1}\theta^w_{m-1})}{\partial \theta^w_{m-1}} \\
&- \sum_{\tilde{y}} p(\tilde{y}|x_t)\lambda_{i,\tilde{y}} \frac{\partial h(\hat{a}_{m-1}\theta^w_{m-1})}{\partial \theta^w_{m-1}}] \\
&= \sum_t^N \sum_i [\lambda_{i,y_t} h'_i(\hat{a}_{m-1}\theta^w_{m-1})\hat{a}_{m-1} \\
&- \sum_{\tilde{y}} p(\tilde{y}|x_t)\lambda_{i,\tilde{y}} h'_i(\hat{a}_{m-1}\theta^w_{m-1})\hat{a}_{m-1}]
\end{aligned}$$

we can decompose the gradient term and derive

$$\delta_{m-1} = \lambda_{i,y_t} h'(\hat{a}_{m-1}\theta^w_{m-1}) - \sum_{\tilde{y}} p(\tilde{y}|x_t)\lambda_{i,\tilde{y}} h'(\hat{a}_{m-1}\theta^w_{m-1}).$$

where DCNF propagates δ_{m-1} to the lower layers so that it can calculate the gradient of these layers. One thing to notice is that the gradient is calculated with $\hat{a}_{m-1}$ instead of a_{m-1} due to the influence of dropout.

5 Experiments

Our main experiment is designed to evaluate the performance of our DCNF model on co-verbal gesture prediction from verbal content and prosody. The following subsection presents our dataset, gesture annotation, input features, baseline models and methodology. To help assess the generalization of our DCNF, we evaluated the performance with a well-studied handwriting recognition (optical character recognition) task [34].

5.1 Co-verbal Gesture Prediction Experiments

The dataset consists of 15 videos which in total represent more than 9 hours of interactions taken from a large-scale study focusing on semi-structured interviews [15]. Our experiment focused on predicting the interviewee's gestures from his/her utterance content and prosody. All the videos were segmented and transcribed using the ELAN tool [3]. Each transcription was reviewed for accuracy by a senior transcriber.

Data segmentation The data is segmented into sequences based on the speaking period. The segmentation can be due to a long pause or the interviewer asked a question. Each frame in the sample data is defined to be 1 second of the conversations. Some of the sequences contained only a very short sentence in which the interviewee replied to the question of the interviewer with a short answer such as "yes/no". We removed all sentences that are less than 3 seconds. The resulting dataset has total 637 sequences with average length of 47.54 seconds.

Gestural sign annotation In the annotation process, we first trained the annotators with the definition of all gestural signs and showed a few examples for each gestural sign. The annotator then used the ELAN tool, looked at the behavior of the participants only when they are speaking, and marked the beginning and the ending time of gestural signs in the video. There will be at most one gestural sign at any time in the data. The annotation results were inspected to analyze the accuracy and insure the annotator had well understood the definition of gestural signs.

Linguistic features Linguistic features encapsulate the utterance content and help determine the corresponding gestures. The extracted data has 5250 unique words, but most of them are unique to a few speakers. To make the data more general, we remove words that happen fewer than 10 times among all the 15 videos, and the resulting number of unique words is down to 817. We represent features as a binary values so that features will be set to 1 when the corresponding linguistic features appear in the corresponding time frame, and 0 otherwise. The linguistic features at the previous time frame and the next time frame are also helpful. In particular, a gesture can for example, proceeds its corresponding linguistic features. Therefore, when a linguistic feature appears at a time frame, its appearance will also be marked in the previous and the next time frame.

The data collection process extracted text from the transcript and also ran a part-of-speech tagger [2] to determine the grammatical role of each word. POS tags are encoded at the word level and are automatically aligned with the speech audio through using the analyzing tools of FaceFX.

Prosodic features In terms of prosody, the data extracted the following audio features: normalized amplitude quotient (NAQ), peak slope, fundamental frequency (f0), energy, energy slope, spectral stationarity [31]. The sampling rate is 100 samples per second. All prosodic features within the same time frame are concatenated into one feature vector. As the time frame is 1 second and the sampling rate is 100 in our dataset, all 100 samples are concatenated into one feature vector as the prosodic features for that time frame. The extraction process also determines whether the speaker is speaking based on f0, and for the periods in the speech that identified as not speaking all audio features are set to zero.

Baseline models Our experiments compared DCNFs with models representing state-of-the-art approaches. We include CRFs, which is applied in the state-of-the-art work [22] on gesture prediction, for comparisons. We also compared with the second-order

CRFs. Additionally, we include support vector machines (SVMs) and random forests, two effective machine learning models. The SVM is an approach that applies kernel techniques to help find better separating hyperplanes in the data for classifications. The random forest is an ensemble approach which learns a set of decision trees with bootstrap aggregating for classification. Both approaches have a good generalization in prior work. Additionally, two existing works that combine CRFs and neural networks, CNF [28] and NeuroCRF [10], are evaluated in the experiment. The experiment also evaluated the performance of DCNFs without using the sequential relation learned from CRFs (denoted as DCNF-no-edge).

Methodology The experiments use the holdout testing method to evaluate the performance of gesture predictions in which the data is separated into training, validation, and testing sets. We trained DCNFs with three hidden layers each with 256 hidden nodes and set the initial learning rate to 0.1 with 0.0003 degrading rate at each iteration. The choice of these hyperparameters are determined based on the validation results. The final result is the performance on the testing set. Each videos in the co-verbal gesture dataset corresponds to a different interviewee. We chose the first 8 interviewees (total clip length correspond to 50.86% of the whole dataset) as the training set, 9 through 12 interviewees (23.18% of the whole dataset) as the validation set, and last 3 interviewees (25.96% of the whole dataset) as the testing set.

Results The results are shown in Table 2. Both the DCNF and DCNF-no-edge models outperform other models. The performance similarity of DCNFs with and without edge features suggest that the major improvement comes from the exploitation of deep architecture. In fact, models that rely mainly on sequential relation show significantly lower performance, suggesting the bottleneck on co-verbal gesture prediction lies in the realization of the complex relation between speech and gesture. The results are unexpected, as based on the work of McNeill, Calbris and others [4, 25], it is reasonable to expect temporal dependencies. Calbris talks of ideation units and rhythmic-semantic units that span multiple gestures, for example. The fact that our models could not exploit temporal dependencies may due to that some of the the gestural signs defined in this task obscure the temporal dependency. For example, some gestural signs that express semantic meanings more specifically can break this kind of temporal correlation. Take wipe as an example, when someone does a wipe, it does not indicate much about whether a frame or a shrug will follow. Given that these are co-speech gestures, if a dependency at this aggregate/abstract level would to occur at the gesture level, it suggest that the same constraint should co-exist at the language level. However, since a speaker can reorder or compose different phrases, it is essentially common for a speaker to alter the verbal content and the underlying gestural behaviors. On the other hand, other subsets of gestural signs might reveal stronger dependencies, for example ones comprising rhetorical structures like enumeration and contrasts, or gestural signs tied to the establishment of a concept such as a container gesture showing a collection of ideas, followed by operations on the concept, such as adding or removing ides/items from the container. Even in these cases, there is the question of whether the features currently being used make it feasible to learn such dependencies. In addition to these fundamental difficulty on formulating the temporal relation, another possible reason is that the data collected in this task may still be too limited for learning the temporal relation.

Models	Accuracy(%)
CRF [22]	27.35
CRF second-order	28.15
SVM	49.17
Random forest	32.21
CNF [28]	48.33
NeuroCRF [10]	48.68
DCNF-no-edge	59.31
DCNF (our approach)	59.74

Table 2: Results of co-verbal gesture prediction.

5.2 Handwriting recognition

To access the generality of DCNFs, we also applied it to a standard hand writing recognitions dataset [34]. This dataset contains a set of (total 6877) handwriting words collected from 150 human subjects with average length of around 8 characters. The prediction targets are lower-case characters, and since the first character is capitalized, all the first characters in the sequences are removed. Each word was segmented into characters and each character is rasterized into 16 by 8 images. We applied 10-fold cross validation (9 folds for training and 1 fold for testing) to evaluate the performance of our DCNF model and compare the results with other models. We trained DCNFs with three hidden layers each with 128 hidden nodes and set initial learning rate to 0.2 with 0.0003 degrading rate at each iteration. The choice of these hyperparameters are also determined based on the validation results

Baseline models In addition to the models compared in the gesture prediction task, this experiment also compared with the state-of-the-art result previously published using the structured prediction cascade (SPC) [37]. The SPC is inspired by the idea of the classifier cascade (for example, boosting) to increase the speed of the structured prediction. The process starts filtering possible states at 0-order and then gradually increase the orders with considering only the remaining states. While the complexity of a conventional graphical model grows exponentially with the order, SPC's pruning approach reduces the complexity significantly and therefore allows applying higher order models. The approach is the state-of-the-art results on the handwriting recognition task. The comparison results of DCNFs with SPC, along with other existing models, are shown in Table 3

Results In this handwriting recognition task DCNF shows improvement over published results. Compared to the gesture prediction task, the mapping from input to prediction

Models	Accuracy(%)
CRF	85.8
CRF second-order	93.32
SVM	86.15
Random forest	96.97
CNF	91.11
NeuroCRF [10]	95.44
DCNF-no-edge	97.21
Structured prediction cascades [37]	98.54
DCNF (our approach)	99.15

Table 3: Results of handwriting recognition. Both the results of NeuroCRF and Structured prediction cascades are adopted from the original reported values.

targets is easier to realize in this task, and therefore the sequential information provides an influential improvement, as shown by the improvement of DCNF over DCNF-no-edge. We have also applied [10, 13] on the task and the results are similar to DCNF-no-edge.

6 Conclusion

Gesture generation presents a novel challenge to machine learning: prediction of gestures must take into account the conceptual content in utterances, physical properties of speech signals and the physical properties of the gestures themselves. To address this challenge, we proposed a gestural sign scheme to facilitate supervised learning and presented the DCNF model, a model to jointly learn deep neural networks and second-order linear chain temporal contingency. Our approach can realize both the mapping relation between speech and gestures and the temporal relation among gestures. Our experiments on human co-verbal dataset shows significant improvement over previous work on gesture prediction. A generalization experiment performed on handwriting recognition also shows that DCNFs outperform the state-of-the-art approaches.

Our framework predict gestural signs from speech, and by combining with existing gesture generation system, for example [9], the overall framework can be applied to animate virtual characters' gestures from speech. The framework relies only on linguistic and prosodic features that could be derived from speech in real-time, thus allowing for real-time gesture generation for virtual character.

7 Acknowledgements

The projects or effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Bergmann, K., Kahl, S., Kopp, S.: Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In: 13th Conference on Intelligent Virtual Agents. pp. 203–216 (2013)
2. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. OReilly Media Inc (2009)
3. Brugman, H., Russel, A., Nijmegen, X.: Annotating multi-media / multimodal resources with elan. In: In Proceedings of the Fourth International Conference on Language Resources and Evaluation. pp. 2065–2068. LREC 2004 (2004)
4. Calbris, G.: Elements of Meaning in Gesture. Gesture Studies 5, John Benjamins, Philadelphia (2011)
5. Cassell, J., Prevost, S.: Distribution of semantic features across speech and gesture by humans and computers. In: Workshop on the Integration of Gesture in Language and Speech (1996)
6. Cassell, J.: Embodied conversational interface agents. Commun. ACM 43(4), 70–78 (Apr 2000)
7. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 477–486. ACM, New York, NY, USA (2001)
8. Chiu, C.C., Marsella, S.: How to train your avatar: A data driven approach to gesture generation. In: 11th Conference on Intelligent Virtual Agents. pp. 127–140 (2011)
9. Chiu, C.C., Marsella, S.: Gesture generation with low-dimensional embeddings. In: Proceedings of the 13th international joint conference on Autonomous agents and multiagent systems. AAMAS '13 (2014)
10. Do, T., Artieres, T.: Neural conditional random fields. In: International Conference on Artificial Intelligence and Statistics (AI-STATS). pp. 177–184 (2010)
11. Ennis, C., McDonnell, R., O'Sullivan, C.: Seeing is believing: body motion dominates in multisensory conversations. In: ACM SIGGRAPH 2010 papers. pp. 91:1–91:9. SIGGRAPH '10, ACM, New York, NY, USA (2010)
12. Ennis, C., O'Sullivan, C.: Perceptually plausible formations for virtual conversers. Computer Animation and Virtual Worlds 23(3-4), 321–329 (2012)
13. Fujii, Y., Yamamoto, K., Nakagawa, S.: Deep-hidden conditional neural fields for continuous phoneme speech recognition. In: International Workshop of Statistical Machine Learning for Speech (IWSML) (2012)
14. Goldin-Meadow, S., Alibali, M.W., Church, R.B.: Transitions in concept acquisition: Using the hand to read the mind. Psychological Review 100(2), 279–297 (Apr 1993)
15. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., Devault, D., Marsella, S., Traum, D., Rizzo, A.S., Morency, L.P.: The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
16. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2013)
17. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. pre-print (2012), 1207.0580v1
18. Kipp, M.: Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Ph.D. thesis, Saarland University (2004)

19. Kopp, S., Bergmann, K.: Individualized gesture production in embodied conversational agents. In: Zacarias, M., Oliveira, J.V. (eds.) Human-Computer Interaction: The Agency Perspective, Studies in Computational Intelligence, vol. 396, pp. 287–301. Springer Berlin Heidelberg (2012)
20. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. pp. 282–289 (2001)
21. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: 6th Conference on Intelligent Virtual Agents. Lecture Notes in Computer Science, vol. 4133, pp. 243–255 (2006)
22. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 papers. pp. 124:1–124:11. SIGGRAPH '10, ACM, New York, NY, USA (2010)
23. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. ACM Trans. Graph. 28, 172:1–172:10 (December 2009), http://doi.acm.org/10.1145/1618452.1618518
24. Marsella, S.C., Xu, Y., Lhommet, M., Feng, A.W., Scherer, S., Shapiro, A.: Virtual character performance from speech. In: Symposium on Computer Animation. Anaheim, CA (Jul 2013)
25. McNeill, D.: So you think gestures are nonverbal? Psychological Review 92(3), 350–371 (Jul 1985)
26. Mohamed, A.r., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. Audio, Speech, and Language Processing, IEEE Transactions on 20(1), 14–22 (Jan 2012)
27. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Trans. Graph. 27(1), 1–24 (2008)
28. Peng, J., Bo, L., Xu, J.: Conditional neural fields. In: NIPS. pp. 1419–1427 (2009)
29. Rickel, J., Johnson, W.L.: Task-oriented collaboration with embodied agents in virtual worlds. In: Embodied conversational agents, pp. 95–122. MIT Press, Cambridge, MA, USA (2000)
30. Salem, M., Rohlfing, K.J., Kopp, S., Joublin, F.: A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In: RO-MAN, 2011 IEEE. pp. 247–252 (July 2011)
31. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech and Language 27(1), 263–287 (2013)
32. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: creating animated conversational characters from recordings of human performance. In: SIGGRAPH '04: ACM SIGGRAPH 2004 Papers. pp. 506–513. ACM, New York, NY, USA (2004)
33. Sutskever, I., Martens, J., Hinton, G.: Generating text with recurrent neural networks. In: ICML (2011)
34. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
35. Taylor, G., Hinton, G.: Factored conditional restricted Boltzmann machines for modeling motion style. In: Bottou, L., Littman, M. (eds.) Proceedings of the 26th International Conference on Machine Learning. pp. 1025–1032. Omnipress, Montreal (June 2009)
36. Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: An overview. Speech Communication 57(0), 209–232 (2014)
37. Weiss, D., Sapp, B., Taskar, B.: Structured prediction cascades. pre-print (2012), 1208.3279v1
38. Yu, D., Deng, L., Wang, S.: Learning in the deep-structured conditional random fields. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications (2009)