Retrieving Target Gestures Toward Speech Driven Animation with Meaningful Behaviors

Najmeh Sadoughi, Carlos Busso
Multimodal Signal Processing (MSP) Laboratory
The University of Texas at Dallas, Richardson TX 75080, USA
Emails: nxs137130@utdallas.edu, busso@utdallas.edu

ABSTRACT

Creating believable behaviors for *conversational agents* (CAs) is a challenging task, given the complex relationship between speech and various nonverbal behaviors. The two main approaches are rule-based systems, which tend to produce behaviors with limited variations compared to natural interactions, and data-driven systems, which tend to ignore the underlying semantic meaning of the message (e.g., gestures without meaning). We envision a hybrid system, acting as the behavior realization layer in rule-based systems, while exploiting the rich variation in natural interactions. Constrained on a given target gesture (e.g., head nod) and speech signal, the system will generate novel realizations learned from the data, capturing the timely relationship between speech and gestures. An important task in this research is identifying multiple examples of the target gestures in the corpus. This paper proposes a data mining framework for detecting gestures of interest in a motion capture database. First, we train One-class *support vector machines* (SVMs) to detect candidate segments conveying the target gesture. Second, we use *dynamic time alignment kernel* (DTAK) to compare the similarity between the examples (i.e., target gesture) and the given segments. We evaluate the approach for five prototypical hand and head gestures showing reasonable performance. These retrieved gestures are then used to train a speech-driven framework based on *dynamic Bayesian networks* (DBNs) to synthesize these target behaviors.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Conversational Agent*

Keywords

Gesture retrieval, conversational agents, speech-driven animations

ICMI'15, November 09-13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820750.

1. INTRODUCTION

Verbal and nonverbal behaviors are an important aspect of human communications. While speech provides verbal communication, people use gestures to clarify, complement and emphasize their intentions and thoughts. Gestures may convey the same meaning as speech, or they can complement and enrich the message [22]. While 90% of gestures occur while speaking [22], nonverbal behaviors can occur while the subject is listening, providing suitable feedback to the speaker (e.g., backchannel communication). *conversational agents* (CAs) aiming to replicate human-like behaviors should carefully model the complex relationship between gestures and speech.

There are two main approaches used to synthesize human-like gestures: rule-based systems, and data-driven methods [11]. Both approaches have advantages and disadvantages. Rule-based systems derive animations relying on semantic analysis of speech, creating appropriate rules to synthesize behaviors that respond to the message [8, 15]. Since rule-based systems do not capture the large variability of gestures and its dependency on speech, these approaches, may result in movements that seem repetitive and not closely synchronized with the prosodic structure in speech [11]. Data-driven approaches learn the behaviors from available recordings, capturing a larger range of variability in the behaviors. Gestures are tightly connected with prosody (energy, intonation, duration) [6]. Therefore, previous studies have used speech prosody features to derive the gestures of the speaker [4,5,10,17,18,20]. While gestures generated by these methods may be perfectly synchronized with speech, they may not appropriately respond to the underlying content in the message (e.g., nodding for negations). In these cases, the CA will convey behaviors without meaning, reducing the role of gestures in the communication. There are studies that combine the rule-based and data-driven approaches to create the behaviors [21, 26, 29].

Kopp et al. [15] proposed a three layer framework for an *embodied conversational agent* (ECA) composed of intent planning, behavior planning, and behavior realization. The last layer generates the behavior, determining the parameters associated with the amplitude and the synchronization points of the behavior in time. We envision a behavior realization system that bridges gap between the rule-based systems and data-driven approaches, exploiting their benefits and overcoming their limitations. For a given message, our proposed model will generate speech-driven gestures that are constrained by specific behaviors associated with the semantic context (e.g., head shakes for negations). We aim to

produce non-repetitive realization of target gestures. A key step toward this approach is generating enough examples for these target gestures to train the models (e.g., head nods, head shakes).

This study provides a flexible framework to retrieve arbitrary number of gestures in the data. We focus on gestures corresponding to movements of head and hands. Starting from few examples of the target prototypical gestures, the framework searches for similar gestures in the database using a two step approach. First, we train a One-class *support vector machines* (SVMs) to effectively reduce the number of candidates for the target gesture. Second, candidate segments are evaluated with a *dynamic time alignment kernel* (DTAK) framework, which estimates the similarity between the target gestures and the available examples. We demonstrate the performance of the system with prototypical behaviors for three hand and two head gestures, achieving precision ranging from 67% to 92%. After retrieving the examples, we analyze the relationship between the detected gestures and the underlying discourse functions of the message (i.e., discourse functions are semantic functions in the message such as *question*, *affirmation*, and *negation*). The study reveals systematic patterns where discourse functions produce characteristic distribution of these prototypical gestures. We also discuss our proposed approach to generate these prototypical gestures using speech-driven framework based on *dynamic Bayesian networks* (DBNs).

2. RELATED WORK

2.1 Rule-Based versus Data-Driven Systems

To design believable CAs, we need to incorporate naturalistic behaviors that replicate the complex relationship between human gestures and speech. These behaviors need to be semantically connected with the message and carefully synchronized with the prosodic structure in speech.

Studies have proposed approaches to synthesize behaviors from data. Among data-driven systems, speech prosodic features are particularly useful to generate gestures [3, 4, 7, 17, 18, 20]. Beat gestures are commonly used to emphasize words according to their role in the message [22]. Given that prosody is also used to fulfill the same task (i.e., emphasizing and parsing the message), it is not surprising to observe a high correlation between speech features and gestures [6]. However, other types of gestures such as iconic and metaphoric gestures are intrinsically related to the message's content. Although prosody features can provide emotional, emphatic and energy related cues to create gestures that are tightly synchronized with speech, additional information is needed to create gestures responding to the underlying content of the message.

There are several studies where the authors have used rules to animate behaviors [1, 8, 15]. These approaches require to define the behaviors, which may result in repetitive behaviors [11]. Furthermore, creating naturalistic synchrony between speech and behaviors is a challenge.

There are studies that have combined these two approaches. Stone et al. [29] designed a system that searches throughout a speech and motion capture dataset for suitable combination of speech and motion units in the corpus conveying similar meaning and planned behaviors. Their system utilizes a dynamic programming scheme which simultaneously solves the selection of speech and motion units. Marsella et al. [21] developed a framework that creates appropriate gestures depending on the content of speech. It uses speech prosody features to capture the emphasis and emotional state of the speaker. Sadoughi et al. [26] proposed a DBN to constrain a speech driven animation based on semantic labels (e.g., *Question* and *Affirmation*.

2.2 Gesture Segmentation and Classification

An important contribution of this study is a flexible system to retrieve segments in the data conveying target gestures. We will use these segments to constrain the behaviors synthesized by our speech-driven animation. In this context, there are several studies on gesture detection, segmentation and classification that are relevant to this study, which we review in this section.

Zhou et al. [32] proposed the *hierarchical aligned cluster analysis* (HACA) algorithm to dynamically segment and cluster motion capture data into movement primitives. HACA combines a dynamic algorithm with DTAK to automatically segment and identify clusters with similar behaviors. DTAK, which we describe in Section 4.3, estimates the similarity between two sequences. Bozkurt et al. [2] used *parallel hidden Markov models* (PHMMs) to find primitives for hand movements. PHMM is an unsupervised framework composed of several branches with equal number of states, each representing a primitive. They used PHMM to simultaneously segment and cluster the motion capture data. In these two approaches the data is segmented into gesture primitives that are automatically generated using unsupervised frameworks. For our problem, these approaches are less effective since we do not have control to define primitives similar to the target gestures.

Studies have formulated this problem as gesture classification. Joshi et al. [12] presented a framework for gesture classification and segmentation. Starting from a vocabulary of predefined gestures, the system trained a random forest model using a database of video and depth map of the joints. During testing, they use a multi-scale sliding window framework for classification, performing a forward search to segment and classify the target gestures.

There are also some studies on gesture detection. Kovar et al. [16] developed a framework that searches a motion capture data for gestures. They use dynamic time alignment to find the distance between segments and each queried gesture. The retrieved samples are the ones with distances below a threshold. They used these samples to expand the variations of the queries, which are used to repeat the search. Nickel et al. [23] presented a system that identifies pointing gestures and the direction of pointing, on stereo camera data. They trained HMMs for each phase of this gesture (onset, hold and offset). They detected the gesture wherever they find three points in time such that the probability of the first point being in the beginning phase and the last point being in the ending phase was higher than the probabilities of these points belonging to the ending and beginning phase, respectively. Wang et al. [31] proposed a framework to detect three gestures used as commands for a smart-board (lining, pointing, and circling). Their framework is based on HMMs, and they identify the gestures if the confidence value for the gesture is higher than a threshold.

2.3 Contribution of this Study

This study proposes a framework to detect prototypical

(a) So-What (b) To-Fro (c) Regress

Figure 1: Examples of the target hand gestures used in this study.

gestures, which are then analyzed as a function of their underlying discourse functions. The contributions of this paper are as follows. First, we design a framework that mines a motion capture database to find similar gestures to arbitrary number of target gestures. Second, we study the relationship between the retrieved gestures and discourse functions. This analysis is tangible, capturing characteristic patterns of typical behaviors produced by the head and hands. Third, we propose a speech-driven framework based on DBN to illustrate how the retrieved samples can be used to synthesize behaviors constrained by the target gesture. This framework can create a link between rule-based and data-driven systems.

3. DATA

This study relies on the MSP-AVATAR corpus [27]. This corpus is rich in nonverbal behaviors and discourse functions. The corpus provides dyadic interactions with improvisations, where the scenarios are carefully designed to create behaviors associated with the following discourse functions: contrast, confirmation/negation, question, uncertainty, suggest, giving orders, warn, inform, large/small, and pronouns (i.e., uttering pronouns such as "I", "you", and "they"). These discourse functions were selected following the work of Poggi et al. [25] and Marsella et al. [21]. This database also includes the labels for discourse functions, providing a suitable resource for the analysis of nonverbal behaviors and their relationship to mid level representation of speech semantics in the form of discourse functions. This corpus comprises video and audio recordings from six actors during four dyadic sessions. In each dyadic recording only one actor was recorded using the motion capture system, providing detailed facial and upper body motion information for four actors. The details of the corpus are described in Sadoughi et al. [27]. This study uses the recordings of one actress, consisting of 22 sessions (66 minutes).

To train and evaluate our approach, we need to first define our behaviors of interest. Since the corpus contains motion capture data from the upper-body joints of the actors, we consider target behaviors from prototypical hand and head gestures. The selection of the target gestures was influenced by the work of Kipp [14], which are listed in Table 1. These gestures are defined as follow:

Head Shake: gesture defined by one or more head yaw rotation.

Head Nod: gesture defined by one or more head pitch rotation.

So-What: gesture defined by outward movement of the hands such that the hands show open palms at the end of the arc (Fig. 1(a)).

To-Fro: gesture defined by the movements of both hands from one side to the other side (Fig. 1(b)).

Regress: gesture defined by the movement of both hands in circles in the direction toward the body (Fig. 1(c)).

Table 1: List of prototypical gestures considered in this study. The set $Samples_{Examples}$ consists of the samples identified in the corpus used to retrieve similar gestures. The set $Samples_{Test\&Dev}$ consists of the samples annotated on the sessions used as development and testing.

Behavior	Region	$\#Samples_{Examples}$	$\#Samples_{Test\&Dev}$
Shake	Head	27	115
Nod	Head	24	138
So-What	Hand	14	21
To-Fro	Hand	27	29
Regress	Hand	26	73

Head gestures are described with yaw, pitch and roll angles (3D vector). Hand gestures are described by the movements of the arm and forearm. We use three degrees of freedom per joint representing its Euler angles, resulting in 12D vector (2 arms $\times$ 2 joints $\times$ 3 angles). In all the experiments we z-normalized all the angles across the whole data.

We used ANVIL [13] to annotate some of the target behaviors in the videos. We split the 22 sessions into two partitions of 19 and 3 sessions. From the first partition, we manually identify between 14 and 27 examples per target behavior, forming our $Samples_{Examples}$ set (Table 1 lists the exact number of examples). To evaluate the performance of the system, we fully annotated three videos for hand and head gestures ($Samples_{Test\&Dev}$ set). Non overlapped sets in this partition are used as develop and test sets. Table 1 lists the actual number of behaviors associated with each target gesture in the $Samples_{Test\&Dev}$ set.

4. FRAMEWORK

This section describes the proposed method to retrieve behaviors in the database that are similar to the few examples provided for training. Figure 2 shows the overall architecture of the system. The key challenge in this data mining task is that gestures can have different durations and be located at any time in the signal. Therefore, we need to jointly solve the problems of segmenting and detecting the gestures. An exhaustive search is not computationally possible, so assumptions are needed to reduce the search space. We implement a temporal reduction to decrease the length of the data and examples. Then, we use multi scale windows to evaluate the presence of the gestures. The behaviors in the windows are compared with the examples using two ap-

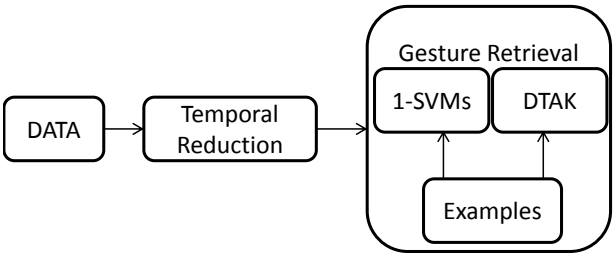

Figure 2: Block diagram of the proposed framework for gesture detection.

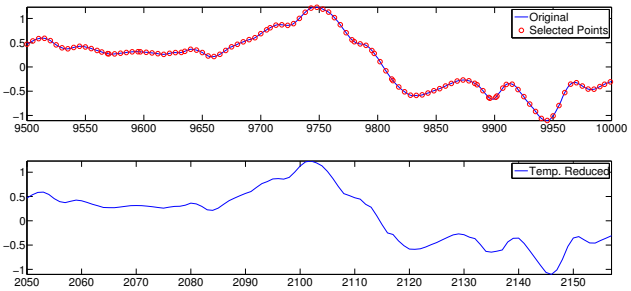

Figure 3: Example comparing the original signal and its temporally reduced version.

proaches. First, we use one class SVMs trained with limited examples. This method discards most of the segments, leaving only candidates segments, which are further evaluated using DTAK. This section explains these steps in details.

4.1 Temporal Reduction

The first step in the process is to simplify the representation of the data, making the search faster. The motion capture data is sampled at 120 fps, where the trajectories convey redundant information. Therefore, it is not practical to run the search algorithm on the original motion capture data. Instead, we remove redundant information from the frames. Zhou et al. [33] proposed an approach for temporal reduction using the K-means algorithm, where the frames were grouped in clusters capturing major transitions in the movements. Then, they applied nonuniform sampling keeping only the transitions between clusters and intermediate frames when the segments were too long. Our proposed approach builds upon this framework. K-means is sensitive to initialization – it can reach local optimum depending on the initial clusters. For each trajectory in the data, we propose to use the *Linde-Buzo-Gray vector quantization* (LBG-VQ) technique [19], setting the number of codebooks to 32. We remove consecutive frames assigned to the same cluster. We keep all the transitioning frames. Furthermore, if more than five consecutive frames are discarded, we keep intermediate frames, following the setting suggested by Zhou et al. [32]. Figure 3 shows an example for temporal reduction showing the same pattern as the original trajectory but with 20% of the frames.

4.2 Gesture Segmentation

The proposed approach simultaneously solves the segmentation and detection of the target gestures. We separately implement the approach for each of the five gestures considered in this study. The segmentation consists of defining multi scale windows, where we search for the gestures. This framework is inspired by the approach proposed by Joshi et al. [12]. We define a window of size L_w frames (after temporal reduction). We set the size of the window two times the longest gesture available in the training data. Within this window, we search for segments using the following approach. First, we define the minimum ($L_{\min}$) and maximum ($L_{\max}$) lengths of the target gesture. Starting from the beginning of the window, we create a segment of length $L_{\min}$, which is incrementally increased by Δ frames until reaching $L_{\max}$ (top diagram in Fig. 4). Then, we repeat the approach, by shifting the starting point of the segments by five frames within the window (middle diagram in Fig. 4). This approach is repeated till reaching the end of the window (bottom of diagram in Fig. 4). At this point, we slide the window keeping 75% overlapped frames, aiming to capture all possible segments containing the target gesture.

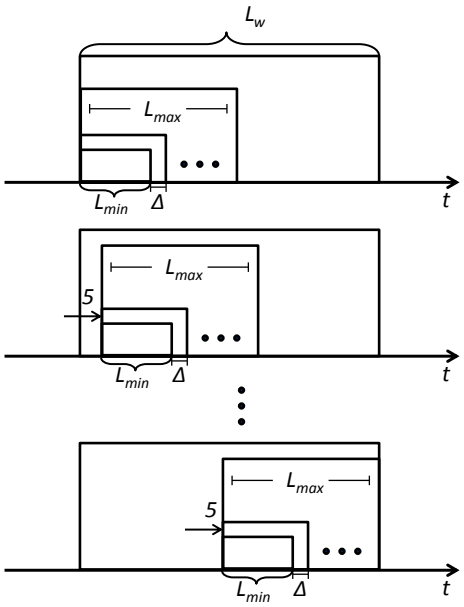

Figure 4: Multi scale window approach for segmentation.

We limit the search space with L_{min} and L_{max}. For a given target gesture, L_{min} equals to 90% of the minimum length of the training samples. Likewise, L_{max} equals to 1.3 times the maximum length of the training samples. The increment frames between iteration is empirically set to $\Delta = (L_{max} - L_{min})/30$. We evaluate the presence of the target gesture in each of the candidate segments using the approach described in Section 4.3. We may detect multiple segments within each window, which may be overlapped. Among the selected gestures in the window, wherever there is an overlap, we select the one with the maximum similarity to the target gesture.

4.3 Gesture Detection

The next step is to determine whether the target gesture is included in the a given segment. Given the large number of segments generated by the multi scale window framework, we implement a two-step approach, where the first step is

fast and efficient to discard irrelevant segments, and the second step is accurate to compare candidate segments with the examples of the target gesture.

The first approach is implemented with one-class SVM with linear kernel [28]. One-class SVM finds a hyperplane that separates the positive samples. This formulation results in a binary function assigning 1 to a small region in the feature space around the positive samples. This framework is ideal for this problem, since it is fast and does not require to define negative samples per gesture. We implement the classifier with LIBSVM [9]. Given the limited number of samples, we train a one-class SVM for each dimension of the feature vector (3 for head gestures, 12 for hand gestures). Instead of using the actual values of the angles, we estimate their standard deviation which are used as features. The standard deviation of the angle captures the dynamic behavior of the gesture. Using only this feature allows us to discard segments without movements or with behaviors that are significantly different from the target gesture. As expected, not all the angles characterizing the target gesture are useful. To select the features, and therefore the classifiers, we use the examples from the training set ($Samples_{Examples}$) using one-sample-out, cross-validation method, where in each fold we train the one-class SVMs with all the samples, excepting the one used for evaluating the performance (notice that samples from the $Samples_{Test\&Dev}$ set are not used). Then, we fused the classifiers using AND operation. If the overall accuracy on the training samples is lower than 85%, we sort the features in descending order according to their overall accuracy. Then, we sequentially remove bad classifiers one by one, till reaching our target 85% accuracy. Notice that these classifiers do not need to have high accuracy. In fact, increasing the accuracy may results in increase of false negative rates, where target gestures are not detected. We can have false positive segments, which can be rejected in the second step.

The second approach is implemented with DTAK, which compares the candidates segments and each of the training examples. DTAK gives a measure of similarity between two segments regardless of their durations [24]. It defines a distance measurement that satisfies the triangular inequality, making it a better fit than *dynamic time warping* (DTW) [32]. Given two sequences X and Y of length l_x and l_y ($X = [x_i, x_2, ..., x_{l_x}]$, $Y = [y_i, y_2, ..., y_{l_y}]$), DTAK computes the similarity between them using the recursive formula given in Equation 1. The variable $u_{i,j}$ is an element of the matrix $U \in \mathbb{R}^{n_{l_x} \times n_{l_y}}$, which contains the accumulated similarity. The first element of the matrix U is initialized as $u_{1,1} = 2K_{1,1}$. $K_{i,j}$ is the kernel capturing the similarity between frames x_i and y_j. The rest of the element in U are found recursively using Equation 1. For DTAK, we use the implementations provided by Zhou et al. [32].

$$\tau(X,Y) = \frac{u_{l_x,l_y}}{l_x + l_y}, \qquad u_{i,j} = \max \begin{cases} u_{i-1,j} + K_{i,j} \\ u_{i-1,j-1} + 2K_{i,j} \\ u_{i,j-1} + K_{i,j} \end{cases} \quad (1)$$

$$K_{i,j} = \exp\left(-\frac{\|x_i - y_j\|^2}{2\sigma^2}\right) \quad (2)$$

This study uses Gaussian kernel (Equation 2). If we set the standard deviation of the kernel close to zero, ($\sigma \approx 0$), the similarity between the two sequences will be low, unless we find an exact match. If σ is too large, it will converge to 1 regardless of the distance between the two sequences. We need to set σ such that it provides a good resolution for this problem. We notice that σ equal to 0.1 gives reasonable performance.

Notice that we have multiple examples per gesture. This framework compares each of the examples with the segment. As a measure of similarity between a segment and the target gesture, Zhou et al. [32] proposed the mean of individual similarities. Instead, we consider the median of the individual similarities obtained across examples (median is less sensitive to outliers than mean). DTAK provides a score with the similarity of the segments and the target gestures. We define a threshold for this metric to make the final decision about a segment. To fix the threshold, we use a three-fold, cross-validation approach on the $Samples_{Test\&Dev}$ data to define the develop (two partitions) and test (one partition) set. We find the threshold that maximizes the F-score on the developing set. To find the F-score, we estimate the precision and recall scores. We consider a true detection, when there is an overlap between the selected segment and the ground truth detection. We consider a false negative when we fail to detect a target gesture. We use the selected thresholds on the corresponding test sets, and evaluate the result. We use the mean of the thresholds found in the three folds to detect gesture in the rest of the data.

Table 2: The precisions and recall of the detected gestures on the test set.

Region	Behavior	19 sessions	$Samples_{Test\&Dev}$	
		Precision [%]	Precision [%]	Recall [%]
Head	Shake	91.32	95.65	42.31
	Nod	85.04	87.10	61.36
Hands	So-What	79.69	76.92	47.62
	To-Fro	59.52	67.86	67.86
	Regress	71.77	78.85	57.75

5. EXPERIMENTAL EVALUATION

We evaluate the proposed approach over the entire data (22 sessions). In total, we retrieve 287 head shakes, 535 head nods, 114 So-What gestures, 223 To-Fro gestures, and 262 Regress gestures. The first evaluation of the proposed approach was to measure the precision over all the sessions, excepting the three sessions used for developing and testing (i.e., 19 sessions). To find the precision, we reviewed the retrieved gestures, to annotate whether they contain the target gesture. When the gestures were correctly detected, we count them as true positives. Otherwise, we count them as false positives. Table 2 shows the results ("19 sessions column). Over 85% of the head gestures successfully convey the target behaviors. While the precision rates decrease for hand gestures, the performance is still over 67%. Head movements are easier to distinguish than hand gestures (3D vectors vs. 12D vectors). Furthermore, there is more confusion given the complexity of hand gestures. In spite of these challenges, these precision rates are very promising.

Evaluating the performance of the system in term of recall rates is more challenging, since we need to account for the gestures that our system fails to identify. For this purpose,

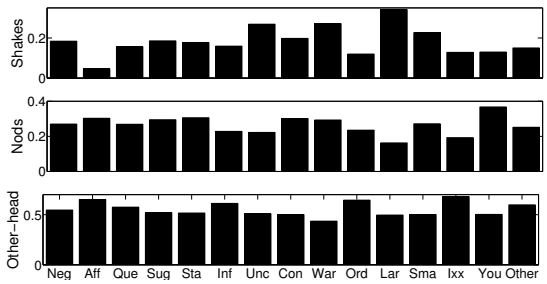

Figure 5: The distribution of the head gestures across discourse functions. Neg: Negation, Aff: Affirmation, Que: Question, Sug: Suggestion, Sta: Statement, Inf: Information, Unc: Uncertainty, Con: Contrast, War: Warn, Ord: Order, Lar: Large, Small: Small, Ixx: I-deictic, You: You-deictic and Other: Other-deictic.

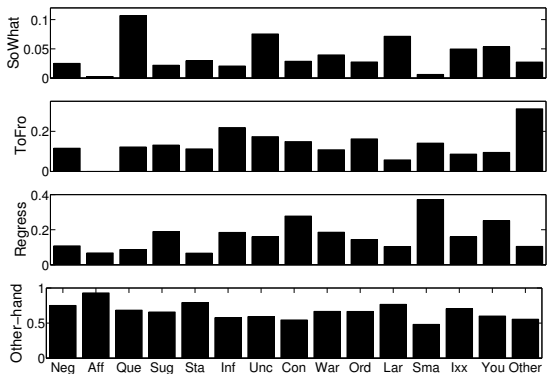

Figure 6: The distribution of the hand gestures across discourse functions. Neg: Negation, Aff: Affirmation, Que: Question, Sug: Suggestion, Sta: Statement, Inf: Information, Unc: Uncertainty, Con: Contrast, War: Warn, Ord: Order, Lar: Large, Small: Small, Ixx: I-deictic, You: You-deictic and Other: Other-deictic.

we use the three test sessions, since they are fully annotated. We consider a false negative when a gesture in the annotation is not detected by our system. Table 2 gives the results for precision and recall rates over test set ($Samples_{Test\&Dev}$ columns). In general, the recall rates are lower than the precision rates indicating that some of the gestures are not detected by the system. For head movements, the recall rates are over 42%, and the precision rates are over 85%. The recall rates are over 57% for *To-Fro* and *Regress* gestures. The precision rates are higher than 65% for all hand gestures

After retrieving the target gestures, we study their distribution along the discourse functions. Figures 5 and 6 show the results for head and hand gestures, respectively. We plot the distribution for each gesture by counting the number of overlapped frames between discourse functions and target gestures. Wherever there is a confusion in a given region (hand or head), we keep the gesture with the highest similarity. We also consider the frames without target gestures ("Other-head", "Other-hand"). These numbers are normalized by the total number of frames associated with a given discourse function (head or hand). Therefore, the sum of the bins corresponding to each discourse function adds to one in both figures.

These results reveal some interesting relationships between these gestures and discourse functions. Figure 5 shows that *Shakes* happen more often during *Large*, *Contrast* and *Warn*. Also, the subject shakes her head more often during *Negations* than during *Affirmation*. We observe that *Nods* happen more often in *You*, *Statement*, *Affirmation* and *Contrast*. In addition, *Nodes* occur more often in *Affirmation* than in *Negation*. For hand gestures, Figure 6 shows that *So-What* happens quite often during *Question*, *Uncertainty* and *Large*. *To-Fro* happens more often when the person is using pronouns referring to *Other* and during *Inform*. *Regress* happens more frequently in *Small*, *Contrast*, *You*, and *Suggest*. These results show that there is a connection between discourse functions and behaviors.

6. SPEECH-DRIVEN ANIMATIONS

We are retrieving gestures to generate speech-drive animations with meaningful behaviors. Our goal is to automatically identify realizations of target gestures in the corpus to synthesize novel versions of these behaviors. After retrieving target gestures from the data, we train a model with these samples to evaluate whether we can capture the characteristics of these behaviors during synthesis. This section describes our approach.

DBNs have been used in other studies to derive animations based on speech prosody features [20, 26]. This study designs a DBN inspired by the constrained DBN proposed in Sadoughi et al. [26]. We propose an improved version of this DBN displayed in Figure 7. The base structure of the model is similar. Ignoring the *Gesture* node, the model consists of a hidden variable representing the state configuration of the model ($H_{h\&s}$). This variable plays the role of a codebook capturing the relationship between prosodic features (*Speech* node) and the target behavior (*Head* or *Hand* node). For a given speech observation, we propagate the evidence through the network, synthesizing the most likely behaviors.

Sadoughi et al. [26] proposed to constrain the models by adding a child node to the state variable $H_{h\&s}$. This node represented the underlying discourse function of the message, which constrained the behaviors (in our work, the constrains are not discourse function, but the actual target behaviors). In this study, we consider the constraint as the cause of the hidden state $H_{h\&s}$ (Figure 7). The target gesture constrains the relationship between speech and gestures. This modification allows the network to learn the prior probabilities of different categories separate from their effects on the variable $H_{h\&s}$. When the number of states is big enough to cover feasible range of behaviors, the characteristics of the constraints can be learned better, even when the data is unbalanced. This network will learn the shared states using the entire data, and the transition probabilities between the states using the constraints. We train the network using full observation, i.e. we have speech prosody features, motion capture data and constraints (i.e., target gesture). To synthesize the behaviors, we run inference on the network using partial observation (just speech and constraints).

We separately train the proposed DBN for head and hand gestures. We consider 16 states for each. The constraints are discrete values which correspond to the presence of a

behavior at each time frame. For a given gesture, we use all the retrieved samples, even if they are incorrect. For head gestures the *Gesture* variable has three states (shake, nod, without-constraint), and for hand gesture it has four states (So-What, To-Fro, Regress, without-constraint). We use the data when the subject is speaking to train the DBN, since we aim to synthesize animations based on speech. Following previous studies [4, 20, 26], speech features are the fundamental frequency (F0) and energy, and their first and second order derivatives. We interpolate the F0 contour to avoid discontinuities in our model. We calculate F0 and energy in windows of 16.67 ms. Since the sampling rate of the motion capture data is 120 Hz, we interpolate the speech features to match the frequency of the motion capture data.

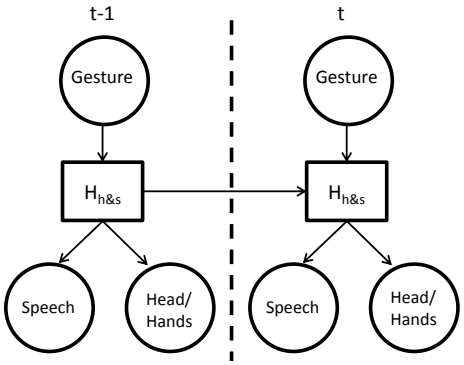

Figure 7: The structure of the DBN which is designed to capture the joint states of speech and movements while constrained on target gestures.

We synthesize animations with the Smartbody toolkit [30]. Using the same sentences, we constrain the DBN with a single behavior at a time, to observe whether the models were able to capture the characteristics of the target gestures. When the person is silent, we create animations without backchannel by simply using the average posture of the person. We smooth the behaviors using the quaternion interpolation of Euler angles used by Busso et al. [5]. We provide a video as a supplemental material with examples of animations driven by speech that are constrained on target gestures by the proposed models. The video shows that the models are in general successful in capturing the target behaviors. When we constrain the animation with head nods or head shakes, the CA displays multiple instances of these gestures. For hand movements, the models were successful in capturing *Regress*. The videos constrained on *So-What* generates more samples of this behavior than during other videos. For *To-Fro*, the approach was less successful. This result may be explained by the limited number of instances retrieved by the system, and the complexity of the behavior.

7. CONCLUSIONS

This paper proposed a framework to automatically detect target gestures defined by few examples in a motion capture database. We considered two head gestures and three hand gestures. The approach jointly solved the segmentation and detection of gestures by using multi scale windows, and a two-step detection framework. The first step efficiently reduced the number of candidate segments using one-class SVM. The second step determined the similarity between segments and the given examples using DTAK. The advantage of this framework is its flexibility to retrieve any gesture. The only requirement is to collect few examples of the target behaviors. We used the retrieved samples to synthesize novel realizations of these gestures using speech-driven animations constrained by these target behaviors. The paper demonstrated that rule-based and data-driven systems can be combined in a principled manner producing gestures with meaning, capturing the variability observed in natural nonverbal behaviors.

There are many opportunities to extend this work. An interesting question is to explore the minimum number of examples per gesture required to achieve acceptable detection rates. Likewise, the framework was evaluated with motion capture data from a single subject in the corpus. We are studying adaptation schemes to generalize the models to retrieve similar gestures from different subjects. We trained a speech driven animation model constrained on the target behaviors. The training data relied on the gestures detected by our framework. The videos show that we were still able to capture the patterns of most of the gesture, even when the detection framework was not perfect. It is not clear how much these errors affect the performance of the speech-driven models. Other statistical models may be more robust against these errors, which can lead to better realization of the behaviors.

8. ACKNOWLEDGMENTS

This work was funded by NSF grant IIS-1352950.

9. REFERENCES

[1] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud. An expressive ECA showing complex emotions. In *Proceedings of the Artificial Intelligence and Simulation of Behaviour (AISB 2007) Annual Convention*, pages 208–216, Newcastle, UK, April 2007.

[2] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and E. Erzin. Multimodal analysis of speech prosody and upper body gestures using hidden semi-Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pages 3652–3656, Vancouver, BC, Canada, May 2013.

[3] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, pages 21–28, New York, NY, USA, 1999.

[4] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.

[5] C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, July 2005.

[6] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, November 2007.

[7] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM*

Transactions on Graphics, 24(4):1283–1302, October 2005.

[8] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, pages 413–420, Orlando, FL,USA, 1994.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27:1–27, April 2011.

[10] C.-C. Chiu and S. Marsella. How to train your avatar: A data driven approach to gesture generation. In *Intelligent Virtual Agents*, pages 127–140, Reykjavik, Iceland, Sep 2011.

[11] M. E. Foster. Comparing rule-based and data-driven selection of facial displays. In *Workshop on Embodied Language Processing, Association for Computational Linguistics*, pages 1–8, Prague, Czech Republic, June 2007.

[12] A. Joshi, C. Monnier, M. Betke, and S. Sclaroffo. A random forest approach to segmenting and classifying gestures. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Sloveni, May 2015.

[13] M. Kipp. ANVIL - a generic annotation tool for multimodal dialogue. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, Aalborg, Denmark, September 2001.

[14] M. Kipp. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, December 2003.

[15] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *International Conference on Intelligent Virtual Agents (IVA 2006)*, pages 205–217, Marina Del Rey, CA, USA, August 2006.

[16] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 559–568, LA,US, Aug 2004.

[17] B. H. Le, X. Ma, and Z. Deng. Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*, 18(11):1902–1914, November 2012.

[18] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics*, 29(4):124:1–124:11, July 2010.

[19] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, Jan 1980.

[20] S. Mariooryad and C. Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2329–2340, October 2012.

[21] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013)*, pages 25–35, Anaheim, CA, USA, July 2013.

[22] D. McNeill. *Hand and Mind: What gestures reveal about thought*. The University of Chicago Press, Chicago, IL, USA, 1992.

[23] K. Nickel and R. Stiefelhagen. Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In *International Conference on Multimodal Interfaces (ICMI 2003)*, pages 140–146, Vancouver, BC, Canada, November 2003.

[24] H. Noma and K. Shimodaira. Dynamic time-alignment kernel in support vector machine. *Advances in neural information processing systems*, 14:921, 2002.

[25] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. de Carolis. Greta. a believable embodied conversational agent. In O. Stock and M. Zancanaro, editors, *Multimodal Intelligent Information Presentation*, Text, Speech and Language Technology, pages 3–25. Springer Netherlands, Dordrecht, The Netherlands, February 2005.

[26] N. Sadoughi, Y. Liu, and C. Busso. Speech-driven animation constrained by appropriate discourse functions. In *International conference on multimodal interaction (ICMI 2014)*, pages 148–155, Istanbul, Turkey, November 2014.

[27] N. Sadoughi, Y. Liu, and C. Busso. MSP-AVATAR corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *1st International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS 2015)*, Ljubljana, Slovenia, May 2015.

[28] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588, 1999.

[29] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513, August 2004.

[30] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, volume 1, pages 151–158, Estoril, Portugal, May 2008.

[31] F. Wang, C.-W. Ngo, and T.-C. Pong. Simulating a smartboard by real-time gesture detection in lecture videos. *IEEE Transactions on Multimedia*, 10(5):926–935, August 2008.

[32] F. Zhou, F. D. la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, March 2013.

[33] F. Zhou, F. Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–7. IEEE, 2008.