GESTURE DYNAMICS MODELING FOR ATTITUDE ANALYSIS USING GRAPH BASED TRANSFORM

Zhaojun Yang, Antonio Ortega and Shrikanth Narayanan

Department of Electrical Engineering, University of Southern California, Los Angeles, CA

ABSTRACT

Gesture dynamic pattern is an essential indicator of emotions or attitudes during human communication. However, there might exist great variability of gesture dynamics among gesture sequences within the same emotion, which form a major obstacle to detect emotion from body motion in general interpersonal interactions. In this paper, we propose a graph-based framework for modeling gesture dynamics towards attitude recognition. We demonstrate that the dynamics derived from a weighted graph based method provide a better separation between distinct emotion classes and maintain less variability within the same emotion class. This helps capture salient dynamic patterns for specific emotions by removing interaction-dependent variations. In this framework, we represent each gesture sequence as an undirected graph of connected gesture units and use the graph-based transform to generate features to describe gesture dynamics. In our experiments, we apply the graph-based dynamics for attitude recognition, i.e., classifying the attitude of an individual as friendly or conflictive. Experimental results verify the effectiveness of our approach.

***Index Terms*—** Attitude, gesture dynamics, graph Fourier transform (GFT), motion capture

1. INTRODUCTION

During human interactions, people unconsciously use their head motions, hand gestures, body postures as well as facial expressions to convey desired communicative messages. Such nonverbal manifestations in human communication are usually triggered and modulated by internal emotional states. Understanding the dynamic patterns presented by nonverbal gesture with respect to underlying emotional states is crucial towards enhancing the development of various applications, such as automatic emotion recognition and human machine interaction.

There is an extensive literature on exploring affective content in gesture dynamics. The features used for describing gesture dynamics are typically based on local spatial position, such as velocities, accelerations or distances [1] [2]. However, since the structure of human gesture is complicated and varying in different time scales and across persons, these specific low-level movement features are not adequate for describing gesture dynamics. Hence higher level tools for gesture dynamics have been developed, and their linkage with emotions or attitudes has been studied. Bernhardt and Robinson have attempted to describe the knocking motion using motion primitives and to detect emotion content from this middle level patterns [3]. More recently, we applied a $1st$ order Markov model to capture the transition structure of hand gesture over a general interpersonal interaction [4].

A major challenge in detecting emotion from gesture in general interpersonal interactions is that body motion could be quite different based on the communicative goals, while the underlying emotion is the same. For example, people with communicative goals, say *to avoid* and *to fight back*, may both express an unfriendly attitude. However, their corresponding gesture dynamics may be quite distinct, i.e., one may be trying to retreat while the other may be moving forward with some intense movements. Moveover, even with the communicative goal that is the same, people may still present different gesture dynamics, since human gesture is varying in different time scales and across persons. This situation can be generalized to time series signals modulated by underlying characteristics. For example, speech signals of the same text content by different speakers could represent different dynamical varabilities. These characteristics urge us to develop methods of modeling time series dynamics that are more robust to the modulation variations.

Graphs provide a flexible representation to model the variations of complex data structure. Graphs have been applied for modeling structured data in multiple applications, such as analysis of social networks [5] or spectral theory for machine learning [6]. Recently, techniques of signal processing on graphs have been developed and applied in many fields [7]. For example, graph-based filtering methods have been developed for denoising images, offering better recognition of texture and edges [8].

In this work, we propose a graph-based framework for modeling the dynamics of a gesture sequence and explore attitude information in the derived gesture variability. The main novelty of our work is representing a gesture sequence as an undirected graph of connected gesture units and applying graph-based transform to describe gesture dynamics. We demonstrate that the derived dynamics from the weighted graph method show a larger divergence across distinct attitudes while maintaining less variability within the same attitude. Our proposed framework could be useful for more general senarios of detecting variability of time series signals modulated differently.

Fig. 1 presents our proposed framework. Given a gesture sequence, we first extract atomic gesture units which define the basic gesture elements [9]. Based on the extracted gesture segments, an undirected graph is constructed for each gesture sequence with each graph node representing a gesture segment. Since the atomic gesture segments contain semantic meaning, such as walking, sitting or crossing arms, constructing graphs with segments rather

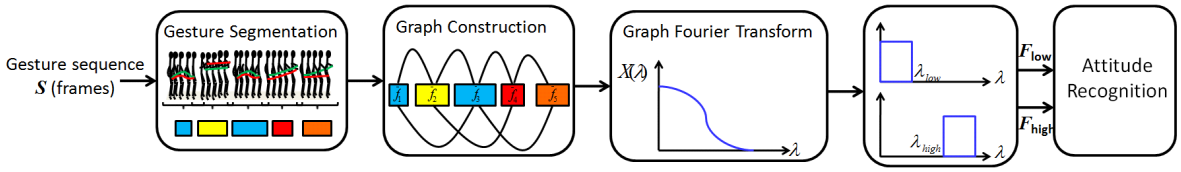

Fig. 1. The proposed graph-based framework for gesture dynamics modeling.

than individual frames can provide a more robust graph structure. Subsequently, we apply the graph Fourier transform to generate features of gesture variability over the corresponding gesture sequence. The graph-based description is then represented in the high and low frequency subbands, describing the oscillation and smoothness of the gesture sequence respectively. The two types of frequency representations are further applied for attitude or emotion recognition. In our experiments, we apply this method on the USC CreativeIT database [10] and achieve an accuracy improvement of 5.7% compared to our recent results in [4].

2. PROBLEM FORMULATION

2.1. Graph Fourier Transform (GFT)

Let $G(V, E)$ be an undirected weighted graph with nodes $V = \{v_i\}_{i=1}^N$ and edges $E = \{e_{i,j}\}_{i,j=1}^N$ between nodes v_i and v_j. Each edge e_{ij} carries a non-negative weight $w_{ij} \geq 0$. If $w_{ij} = 0$, then the two nodes v_i and v_j are not connected. The adjacency matrix $\mathbf{A}$ of the graph G is defined as: $\mathbf{A}(i,j) = w_{ij}$. For an undirected graph where $w_{ij} = w_{ji}$, $\mathbf{A}$ is symmetric. The graph Laplacian is then defined as: $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal degree matrix whose diagonal element d_i is the sum of weights on the edges connected to the node v_i, i.e., $d_i = \sum_{j=1}^N w_{ij}$. In this work, we use the symmetric normalized Laplacian $\mathcal{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$, since the scale of its eigenvalues is invariant to graph structures, which is described below.

Noting that the Laplacian matrix $\mathcal{L}$ is real, symmetric and positive semi-definite, it has a set of orthonormal eigenvectors $\{u_n\}_{n=0}^{N-1}$ associated with real, non-negative eigenvalues $\{\lambda_n\}_{n=0}^{N-1}$, where $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \cdots \lambda_{N-1} \leq 2$. $\lambda_{N-1} = 2$ when the graph G is bipartite. Given a graph G, each graph node is associated with a sample value or vector, and the collection of these samples from all the nodes refers to a graph signal. The graph Fourier transform $\mathbf{X}$ of a graph signal $\mathbf{x}$ on G is defined as the projection of $\mathbf{x}$ onto the eigenvectors of the graph Laplacian:

$$\mathbf{X}(\lambda_n) = u_n^t \mathbf{x}. \tag{1}$$

Analogously to the classic Fourier analysis, the graph Laplacian eigenvectors and eigenvalues also have a frequency interpretation. Eigenvectors corresponding to smaller eigenvalues vary slowly across the graph whereas those associated with larger eigenvalues change rapidly. Hence, the eigenvalues $\{\lambda_0, \lambda_1, \cdots, \lambda_{N-1}\}$ are arranged in a frequency ascending order. Accordingly, the graph Fourier transform $\mathbf{X}(\lambda_n)$ of a graph signal $\mathbf{x}$ reflects the overall fluctuation or dynamics of $\mathbf{x}$ across the graph. A higher energy in the lower frequencies indicates the smoothness of $\mathbf{x}$ across the graph, i.e., the signal values of two connected nodes are more likely to be similar to each other. In contrast, a higher energy in the high frequencies shows a large variability of $\mathbf{x}$. For more details regarding GFT, we refer the readers to [7].

2.2. Graph Construction

Consider a gesture sequence $\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_T]$, where $\mathbf{f}_t$ is a gesture feature vector at frame t, e.g., joint angles or spatial positions of different body parts, and T is the length of the sequence. To model the gesture dynamics of $\mathbf{S}$, we first divide the gesture feature sequence $\mathbf{S}$ into short segments $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_N\}$ using the parallel HMM model in [11], where ϵ_i is the i-th segment described as $[\mathbf{f}_{\epsilon_i^1}, \cdots, \mathbf{f}_{\epsilon_i^{T_i}}]$ and T_i is the length of ϵ_i. This HMM model automatically partitions gesture streams into segments, each of which is assigned into one of C clusters by maximizing the likelihood through Viterbi decoding. We employ the same clustering method as in [4], so that the experimental results are comparable. The segment samples in each cluster are represented by the cluster center $\mathbf{c}_l$, i.e., the mean gesture feature vector in the lth cluster C_l. The segmentation process is illustrated in Fig. 2.

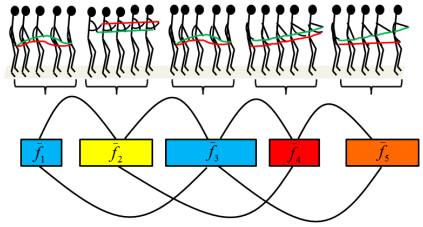

Fig. 2. Illustration of the segmentation process. The red and green lines are the trajectories of right and left hands respectively. The color of each segment indicates the cluster type. Connected segments form the graph structure with $k = 2$.

Based on the extracted segments, we further construct an undirected graph G for each gesture sequence. The graph node v_i represents the gesture segment ϵ_i. The derived segments define the basic gesture units and contain semantic meaning, e.g., walking, sitting or crossing arms. Constructing a graph using segments rather than individual frames can lead to a more robust graph structure. The graph architecture is defined by connecting each gesture segment with its next k neighbors. An example graph with $k = 2$ is shown in Fig. 2. The sample vector associated with the node v_i is defined as $\mathbf{c}_{l^{\epsilon_i}}$, i.e., the center of the cluster C_l that the corresponding segment ϵ_i belongs to. The corresponding graph signal is: $\mathbf{x} = [\mathbf{c}_{l^{\epsilon_1}}, \cdots, \mathbf{c}_{l^{\epsilon_N}}]$. The

edge weights are defined via the Gaussian similarity function: $w_{ij} = exp(-||\mathbf{x}(i) - \mathbf{x}(j)||^2/2\sigma^2)$. Gaussian similarity is a popular graph weight measure and has been widely used for bilateral filtering and machine learning problems [6] [12]. If a spike appears in a gesture sequence, the two dissimilar neighbor segments are connected with a small edge weight. The high frequency energy induced by this spike is thus reduced, and the effect of such significant difference between the two neighbors is alleviated. We assume that there exist certain dynamic patterns with small variations for specific emotion, and such larger variations between segments are more likely to be induced by specific interactions. For instance, there may be more spikes appearing in the gesture sequences of people with *to fight back* goal compared to people with *to avoid* goal. Therefore, weighted graphs may help to capture salient such dynamic patterns by reducing the effect of interaction-dependent variations. We will demonstrate this point in Section 4.

2.3. Gesture Dynamics Representation

Now each gesture sequence $\mathbf{S}_m$ has been transformed into an undirected graph G_m, along with a graph signal $\mathbf{x}_m$ associated to it. According to Section 2.1, the GFT of a graph signal describes its variability across the graph. Hence, the dynamics of a gesture sequence could be represented by $\mathbf{X}_m$ through Eq. (1). It is noteworthy that the graph structures of different gesture sequences differ in terms of edge weights and number of nodes, leading to distinct graph Fourier transform bases, i.e., different sets of graph Laplacian eigenvectors. Thus, graph frequency representations $\mathbf{X}_m$ for different gesture sequences cannot be directly compared to each other. To get a more unified dynamic representation across different graph structures, we propose to partition the graph frequency band into low and high frequency subbands and extract statistical functionals of $\mathbf{X}_m$ respectively from each subband. Specifically, we divide the frequency band $\{\lambda_n\}_{n=1}^{N} \in [0, 2]$ into subbands of low frequency $\{\lambda_{n_{low}}\}_{n_{low}=0}^{N_{low}} \in [0, \tau_{low}]$ and high frequency $\{\lambda_{n_{high}}\}_{n_{high}=N_{high}}^{N-1} \in [\tau_{high}, 2]$. From each subband, we further extract six statistical functionals — min, max, range, std, mean, energy — of $\mathbf{X}_m$ for each gesture feature dimension, denoted as $\mathbf{F}_m^{low}$ and $\mathbf{F}_m^{high}$. $\mathbf{F}_m^{low}$ describes the smoothness of the gesture sequence while $\mathbf{F}_m^{high}$ quantifies its oscillations over an interaction. The two types of dynamic representations will be applied for attitude/emotion recognition.

3. EXPERIMENTS

3.1. Database

In the experiment, we apply our framework on the USC CreativeIT database, which is a multimodal database of dyadic theatrical improvisations [10]. It contains detailed full body Motion Capture (MoCap) data of participants during dyadic interactions, as shown in Fig. 3(a). We manually mapped the 3D locations to joint angles of different body parts using MotionBuilder [13]. Fig. 3(b) illustrates the Euler angles of the arm and forearm in the x, y and z directions. In the experiment, we focus on analyzing hand gesture dynamics. The joint angles of left arm, left forearm, right arm and right forearm are used as hand gesture features $\mathbf{f}$.

(a) Motion Capture Markers. (b) Joint angles for the hand.

Fig. 3. (a) The positions of the Motion Capture markers; (b) The illustration of joint angles for the hand.

The interactions performed by pairs of actors are goal-driven; actors have predefined communicative goals, e.g., to comfort or to avoid. The goal pair of each dyad defines the attitudes of the interlocutors towards each other and the content of the interaction. As defined by the goals, the attitudes of interacting participants can be naturally grouped into classes of friendliness and conflict which will be analyzed in the experiments that follow. There are 46 interactions performed by 16 distinct actors (9 female), including 50 friendly and 42 conflictive individuals.

Since the interactions are improvised, and thus more similar to general interpersonal interactions, there might exist great variability among gesture sequences even within the same attitude class, as introduced in Section 1. After partitioning the gesture sequences into short segments, we find that the number of segments of each sequence ranges from 40 to 128 for the friendly attitude, and from 25 to 63 for the conflictive attitude. The mean velocity of each gesture sequence changes from 127 to 365.3 for the friendly attitude, and from 194.8 to 371.9 for the conflictive attitude. We demonstrate in Section 4 that our proposed framework based on weighted graphs can help to reduce the effect of such interaction-dependent variations and maintain less variability within the same emotion class.

3.2. Experiment setup

To test the effectiveness of our graph-based gesture dynamics for attitude expressiveness, we apply the dynamic features for attitude recognition, i.e., to classify the interaction attitude of the corresponding individual as friendly or conflictive.

In the experiment, we employ support vector machine (SVM) and the leave-one-sample-out scheme. We compare the performance of the unweighted graph and the weighted graph with edge weights defined as Gaussian similarities (see Section 2.2). We also investigate the performance when replacing GFT in our system by the classical DFT and DCT transforms. Note that the GFT of a circulant unweighted graph is equivalent to DFT. The unigram and bigram features capturing the transition dynamics in [4] modeled by Markov model are also used as baseline features. For simplicity, we let the cutoff frequencies τ_{low} and τ_{high} (see Section 2.3) be symmetric with respect to 1, i.e., $\frac{\tau_{high}+\tau_{low}}{2} = 1$ and denote τ_{low} as τ.

3.3. Experimental results

We first investigate how the cutoff frequency τ affects the attitude expressiveness of the graph-based dynamics. The weighted graph is used here. Fig. 4 presents the recognition rates in relation to τ respectively for the low frequency representation F_{low},

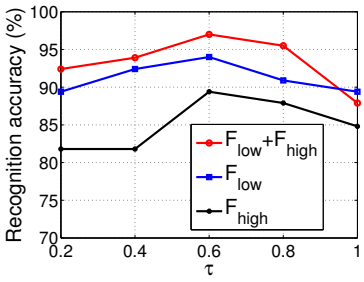

Fig. 4. Recognition accuracy varying with τ.

high frequency representation F_{high} and their combination. τ changes from 0.2 to 1 with a step of 0.2. We can observe that the low frequency representation outperforms the high frequency one. The combination of the two types of dynamic features can further increase the discriminative power in general. The highest performance in each of the three cases is achieved when $\tau = 0.6$.

Table 1 presents the best recognition results achieved for different transform methods by tuning τ, and for the unigram and bigram features in [4]. Overall, the recognition rates of the transform-based dynamics exceed those of the unigram and bigram features, suggesting the usefulness of the transform based features for describing the smoothness and oscillation of gesture sequences. Specifically, we can observe that the best performance of the unweighted graph is the same as that using DFT and DCT, and the weighted graph achieves the highest performance of 97% among all the methods, showing that weights better capture the gesture evolution structure and bring benefits for dynamics modeling.

Table 1. Summary of recognition accuracies (%); the chance rate is 54%.

Transform	F_{low}	F_{high}	F_{low} + F_{high}
Weighted	94	89.4	97
Unweighted	90.9	87.9	92.4
DCT	90.9	92.4	90.9
DFT	92.4	72.7	89.4
Markov [4]	Unigram	Bigram	Unigram + Bigram
	87.8	78.4	91.3

4. ANALYSIS AND DISCUSSION

To better understand the characteristics of gesture dynamics the weighted graph based method captures, we study the cumulative energy function for each gesture sequence using weighted and unweighted graphs respectively. The cumulative energy function describes the proportion of energy that a signal takes up in the frequency region from 0 to λ_k. Given a graph Fourier transform $\mathbf{X}$, the energy sequence for the d-th dimension is $E_d = [|X_d(\lambda_0)|, \cdots, |X_d(\lambda_{N-1})|]$. E_d is further normalized as $\bar{E}_d$ such that $||\bar{E}_d||_2 = 1$. The cumulative energy at λ_k for the d-th dimension is: $F_E(\lambda_k) = \sum_{i=0}^{k} \bar{E}_d(\lambda_i)$.

For better presentation, we keep only the $1st$ PCA component of the gesture feature vectors $\mathbf{f}$, which retains 58.1% of total variance, and then apply GFT on the reduced graph signals. Fig. 5 presents the sample mean cumulative energy curves of the first PCA component within each attitude class using weighted and unweighted graphs respectively. In general, the energy profiles of friendly and conflictive attitudes using weighted graphs can be distinguished from each other, while only a small difference can be observed between the two attitudes using unweighted graphs. We then apply GFT to the original feature vector without using PCA, and calculate the difference between the mean cumulative energies of the two types of attitudes for each dimension. The mean difference between friendly and conflictive attitudes across dimensions and across frequencies is 0.038 for weighted graphs and 0.02 for unweighted graphs. These results imply that the dynamics derived using weighted graphs provide a better separation between emotions.

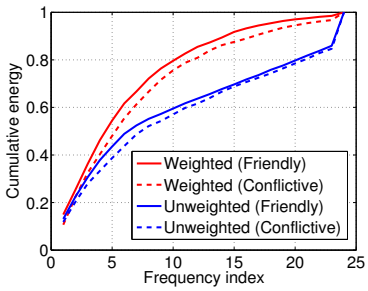

Fig. 5. Sample mean cumulative energy curves of the first PCA component using weighted and unweighted graphs for different attitudes.

In addition, we investigate the variance of energy profiles within an attitude class for weighted and unweighted graphs. The mean variance across dimensions within the friendly attitude using weighted graphs is significantly smaller than using unweighted graphs with $p \ll 0.0001$. Similar results are obtained for the conflictive attitude with $p = 0.0085$. Hence, the weighted graph based dynamics have less variability across gesture sequences within the same emotion class while keeping a larger divergence across distinct emotion classes. We can also observe from Fig. 5 that the high-frequency energy is reduced when using weighted graphs. This indicates that the weighted graph based method might help to remove the interaction-dependent gesture variations, and could better capture the salient dynamic patterns with respect to specific emotion.

5. CONCLUSIONS

Gesture dynamic pattern is an essential indicator of emotions or attitudes during human communication. In this paper, we proposed a graph-based framework for modeling dynamics of a gesture sequence towards attitude recognition. The main novelty of this work is representing a gesture sequence as an undirected graph and using graph-based transform to generate features to describe gesture dynamics. We demonstrate that the dynamics derived from the weighted graph method provide a better separation between distinct emotions and maintain less variability within the same emotion class. Experimental results verify the effectiveness of our approach. In the future, it would be interesting to derive gesture dynamics in multi-resolution, and apply this framework to other published emotion databases.

6. REFERENCES

[1] A. Kapur, V-B. Naznin, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *Affective Computing and Intelligent Interaction*, pp. 1–7. Springer, 2005.

[2] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing, Special Issue on Continuous Affect Analysis*, 2012.

[3] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Affective Computing and Intelligent Interaction*, pp. 59–70. Springer, 2007.

[4] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan, "Analysis of interaction attitudes using data-driven hand gesture phrases," in *Proc. of ICASSP*, 2014.

[5] M.O. Jackson, *Social and economic networks*, Princeton University Press, 2010.

[6] M. Szummer and T. Jaakkola, "Partially labeled classification with markov random walks," in *nips*, 2001, pp. 945–952.

[7] D.I. Shuman, S.K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 83–98, 2013.

[8] F. Zhang and E. R. Hancock, "Graph spectral image smoothing using the heat kernel," *Pattern Recognition*, vol. 41, no. 11, pp. 3328–3342, 2008.

[9] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," *The relationship of verbal and nonverbal communication*, vol. 25, pp. 207–227, 1980.

[10] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Proc. of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, 2010.

[11] M.E. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, 2008.

[12] Akshay Gadde, Sunil K Narang, and Antonio Ortega, "Bilateral filter: Graph spectral interpretation and extensions," in *ICIP*, 2013.

[13] Installation Guide, "Autodesk®," 2008.