Weighted Geodesic Flow Kernel for Interpersonal Mutual Influence Modeling and Emotion Recognition in Dyadic Interactions

Zhaojun Yang[1], Boqing Gong[2], and Shrikanth Narayanan[1]
[1]*Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA*
[2]*Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816 USA*
Email: zhaojuny@usc.edu, bgong@crcv.ucf.edu, shri@usc.edu

***Abstract*—Interpersonal mutual influence occurs naturally in social interactions through various behavioral aspects of spoken words, speech prosody, body gestures and so on. Such interpersonal behavior dynamic flow along an interaction is often modulated by the underlying emotional states. This work focuses on modeling how a participant in a dyadic interaction adapts his/her behavior to the multimodal behavior of the interlocutor, to express the emotions. We propose a weighted geodesic flow kernel (WGFK) to capture the complex interpersonal relationship in the expressive human interactions. In our framework, we parameterize the interaction between two partners using WGFK in a Grassmann manifold by fine-grained modeling of the varying contributions in the behavior subspaces of interaction partners. We verify the effectiveness of the WGFK-based interaction modeling in multimodal emotion recognition tasks drawn from dyadic interactions.**

1. Introduction

Interpersonal mutual influence occurs naturally in social interactions, notably in the unfolding emotional dynamics, through various behavioral aspects, such as spoken words, speech prosody, body gestures. Indeed, in order to accomplish effective communication, individuals usually adapt their verbal and non-verbal behaviors to those of their interaction partners. Such behavior adaptation (also related to variedly known notions of entrainment and coordination) is synchronized in time and expressed as either similar or dissimilar behaviors [1]. The mutual behavior effect controls the dynamic flow of a conversation, shapes the overall interaction patterns, and facilitates the communication to move smoothly, efficiently, and coherently.

Human communication is an interactive process of continuously unfolding human behaviors, established on a common ground of interaction participants. Emotions, a major component of the communication structure, play a crucial role in how we think and behave and allow humans to understand each other better. The underlying emotional states modulate not only the multimodal behavior dynamics of individuals but also the interpersonal behavior dynamic flow. Understanding and computationally modeling the underlying emotional effect on interaction dynamics of human behavior can bring insights into automating emotion recognition and facilitate the design of intelligent human-machine interfaces in a variety of applications.

This work focuses on exploring how an interacting participant adapts his/her behavior to the multimodal language, i.e., gesture and speech, of the interlocutor, to express the internal emotions in a dyadic interaction. In particular, we propose a weighted geodesic flow kernel (WGFK) to capture the mutual influence between dyadic interaction partners. The geodesic flow kernel (GFK) has been previously used to solve domain adaptation problems [2]; such as for achieving good recognition performance on mobile phone images (target domain) using the classifiers trained on Web images (source domain). Yang and Narayanan, instead, used GFK to parameterize the interaction between two partners and showed superior results on emotion recognition in dyadic interactions [3].

However, GFK simply averages the interaction cues embedded in a series of subspaces in a Grassmann manifold, without fine-grained modeling of their potentially varying contributions to the interpersonal mutual influence. We argue that it is suboptimal to use GFK in such a way, and note that the interacting partners could possess distinct behavior idiosyncrasies and thus should be modeled as two domains. Human communication involves a variety of dynamics and complexity over time, making it difficult to fully capture such complex interpersonal relationship in real-life interactions. Such dynamics may be embedded in different interaction modality subspaces to different degrees, but GFK treats them equally important. To this end, we propose to improve GFK [2], [3] by weighing the subspaces along the geodesic flow according to the detailed expression interactions. This effectively increases GFK's modeling capacity, enabling the resultant approach, which we call weighted GFK (WGFK), to better capture the interpersonal interaction characteristics. In what follows, we discuss related work, present the philosophy behind choosing the weight function for WGFK, and then verify its effectiveness in multimodal emotion recognition.

2. Related Work

The behavior adaptation phenomenon in human communication in terms of vocal patterns, head motion, and body gestures has been well-established in the psychology

domain. For example, in the research on interpersonal relations, behavior synchrony in a couples interaction has been shown to offer predictive markers of the couples mental distress and well-being conditions [4] [5]. Chartrand *et al*. described that humans unconsciously mimic the behavior of their interaction partners to achieve more effective and pleasant interactions [6]. Ekman found that body language of interviewees is distinctly different between friendly and hostile job interviews [7]. Neumann *et al*. reported that the emotions in speech would induce a consistent mood state in the listeners [8]. Kendon qualitatively described detailed interrelations between movements of the speaker and the listener by analyzing sound films of social interactions [9]. He also found that the movement of the listener might be rhythmically coordinated with the speech and movement of the speaker in a social interaction. The work in [10] has demonstrated a high degree of unintentional coordination between rhythmic limb movements of two partners.

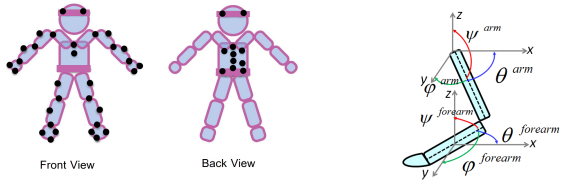
Many engineering works have also been developed based on this mutual influence of interaction subjects. Levitan *et al*. found that interacting partners tend to utilize similar sets of backchannel-preceding cues which are a combination of speech cues of an individual in response to one's interlocutor [11]. Morency *et al*. predicted head nods for virtual agents from the audio-visual information of a human speaker based on sequential probabilistic models [12]. Heylen *et al*. studied what types of appropriate responses, e.g., facial expressions, an agent should display when a human user is speaking, to increase rapport in human-agent conversation [13]. Researchers have also used the emotional state of an interlocutor to inform that of a speaker by modeling emotional dynamics between two interaction participants [14] [15]. The influence framework proposed in [16] models participants in conversational settings as interacting Markov chains. Lee *et al*. proposed prosody-based computational entrainment measures to assess the coordination in the interactions of married couples [17]. Robotics research has shown that human subjects use the robot's cues to regulate conversations and to convey affective intentions, resulting in a smoother interaction with fewer interruptions [18].

Since emotion is one of the major elements influencing the multimodal channels of human speech, body gestures, and facial expressions, the interaction patterns of a dyad's behavior are accordingly shaped by the underlying emotional states [1]. For example, two participants with friendly attitudes may tend to approach each other, while those with conflictive attitudes may try to fight with or avoid each other. The analysis work in [19] has empirically revealed that the coordination patterns of a dyad's behavior depend on the interaction stances assumed (e.g., friendly *vs.* conflictive). Lee *et al*. also investigated the relationship between affective states (positive *vs*. negative) and the vocal entrainment strength in married couples' interactions. A higher degree of vocal entrainment was found for couples with positive attitudes [20].

Motivated by these findings on the interrelation between emotions and interpersonal influence, researchers have made progress in modeling such mutual effect in interactions for enhancing the performance of recognizing the emotional states of individuals [15] [21]. The benefits of incorporating mutual influence into the emotion recognition framework have been validated in these studies. There has also been much research on modeling the behavioral interaction for action recognition [22] [23]. Most of these studies have relied on training with statistical models, e.g., coupled HMMs, which usually require significant amounts of data. Mariooryad *et al*. exploited emotion-related patterns in behavioral interactions [24], which is similar to this work. However, they simply concatenated behavioral information of two partners for assessing the emotional state of an individual without modeling the latent interaction structure of the dyad's behavior.

3. Database Description

In this work, we use the USC CreativeIT database for dyadic interaction modeling [25] [26]. It is a multimodal database of dyadic theatrical improvisations performed by pairs of actors. Interactions are goal-driven; actors have predefined goals, e.g., *to comfort* or *to avoid*, which can elicit natural realization of emotions as well as expressive multimodal behavior. There are 50 interactions in total performed by 16 actors (9 female). The audio data of each actor was collected through close-talking microphones at 48 kHZ. A Vicon motion capture system with 12 cameras captured the detailed full body Motion Capture (MoCap) data at 60 fps, i.e., the (x, y, z) positions of the 45 markers of each actor, as shown in Fig. 1(a).



(a) Motion Capture Markers. (b) Angles for hand joints.

Figure 1. (a) The positions of the Motion Capture markers; (b) The illustration of Euler angles for hand joints.

3.1. Gesture and Acoustic Features

This work focuses on multimodal behavior of speech and hand gesture which are highly expressive forms in human communication. We manually mapped the motion data, i.e., the 3D locations of markers, to the angles of different human body joints using MotionBuilder [27]. The joint angles are popular for motion animation [28] [29] and have also been applied for exploring attitude-related gesture dynamics in our previous work [30]. Fig. 1(b) illustrates the Euler angles (θ, ϕ, ψ) of hand joints (arm and forearm) in x, y, z directions. The angles of both right and left hand joints are used as hand gesture features. In addition, we extracted acoustic features of pitch and the rms energy, as well as 12 Mel Frequency Cepstral Coefficients (MFCCs)

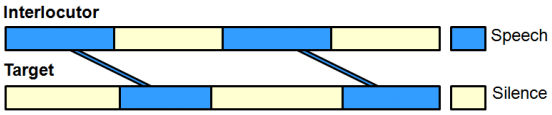


Figure 2. Illustration of setting up dialog turn pairs. The target dialog turns are with emotion annotations.

for each actor. These features were extracted every 16.67 ms (60 fps) with an analysis window length of 30 ms, in order to match with the MoCap frame rate. The pitch features were smoothed and interpolated over the unvoiced/silence regions. We further augment both hand gesture and acoustic features with their $1st$ derivatives to incorporate the temporal dynamics.

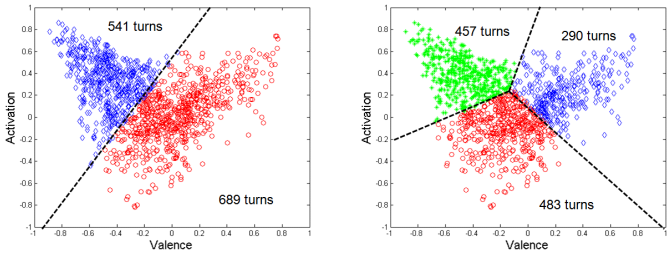
3.2. Emotion Labels

The emotional state of each actor was annotated in terms of activation (excited *vs*. calm) and valence (positive *vs*. negative) by three or four annotators. To capture the continuous flow of body gestures during an improvisation, we annotated time-continuous emotion for each actor throughout the interaction. Annotators used the Feeltrace instrument [31] to time-continuously indicate the emotion attribute value from -1 to 1 for each actor while watching the video recording. More details of the annotation process can be found in [32].

As described in [32], we define the inter-rater agreement for the continuous emotion annotations as the linear correlation between two annotators. For each actor recording, we compute the correlation between every pair of annotators and only keep the annotator pairs with correlations greater than 0.5. We further partition each actor recording into dialog turns according to speech regions. As a result, we have 1230 annotated dialog turns (referred to as the target turn hereafter) in total. Each target turn is paired with the corresponding interlocutor's previous turn, as illustrated in Fig. 2. Our work focuses on modeling interaction behavior between the paired dialog turns. The values of activation and valence of each target turn are calculated by averaging the annotations among frames and across annotators. We jointly consider activation and valence by creating K emotional clusters in the valence-activation space using k-means algorithm. Such K-class recognition scheme has also been adopted in [24] [33]. We consider clusters with $K = 2$ and $K = 3$, and Fig. 3 shows the corresponding clustering results.

4. Approach

In this section, we elaborate our approach to the modeling of interpersonal mutual influence in affective dyadic interactions. We first revisit the geodesic flow based modeling and then show how to improve it by a tailored weight function. We verify the effectiveness of the proposed weighted geodesic flow kernel (WGFK) on the emotion recognition task from the affective dyadic interactions.



(a) $K = 2$. (b) $K = 3$.

Figure 3. Resulting emotion classes in the valence-activation space for $K = 2$ and $K = 3$.

4.1. Revisiting the Geodesic Flow Based Interaction Modeling

The main objective of this work is to model the mutual influence of a dyad in an affective interaction. Seeing the partners involved in the interaction as two "domains", Yang and Narayanan propose to model them based on the geodesic flow on a Grassmann manifold, originally developed for domain adaptation [2].

Specifically, each dialog turn of a target participant is paired with the corresponding interlocutor's previous turn, as illustrated in Fig. 2. Such turns are related by the participants' interactions and yet are also distinct from each other, due to different personal behavior habits. Therefore, if we respectively embed the turns of the two participants into two subspaces $\mathbf{P}_T \in \mathcal{R}^{D\times d}$ and $\mathbf{P}_I \in \mathcal{R}^{D\times d}$ of lower-dimensions than the dimensionality D of their feature representations $\mathbf{x} \in \mathcal{R}^D$, we expect the two subspaces to overlap to some extent. Geometrically, any two subspaces could be connected continuously by a shortest path on the Grassmann manifold, along which each point itself is also a subspace of the same dimension. The shortest path is called a geodesic flow $\mathbf{\Phi}(t)$, where t is between 0 and 1, $\mathbf{\Phi}(0) = \mathbf{P}_T$, and $\mathbf{\Phi}(1) = \mathbf{P}_I$. We refer the readers to [2] and [3] for the details of computing the geodesic flow.

In the context of this paper, the geodesic flow assembles the behavior characteristics of each participant in the interaction and parameterizes the gradual adaptation from the interlocutor to the target participant. Given two D-dimensional behavior vectors $\mathbf{x}_T$ and $\mathbf{x}_I$ from the target subject and the interlocutor, respectively, their projections into any subspace $\mathbf{\Phi}(t)$ on the flow are calculated by $\mathbf{\Phi}(t)^T\mathbf{x}_T$ and $\mathbf{\Phi}(t)^T\mathbf{x}_I$. Each $t \in [0, 1]$ enforces some particular characteristics of the interaction and suppresses the others. In other words, the infinite number of subspaces $\mathbf{\Phi}(t), t \in [0, 1]$, disentangle the interaction patterns to different aspects. Jointly, all the projections integrate the behavioral characteristics of both interaction partners.

Interaction Matrix. Although the geodesic flow quantitatively represents the mutual influence between the interaction partners, it is computationally infeasible to handle the infinite number of subspaces or projections. To tackle this

problem, an interaction matrix is derived in [3] as follows. Concatenating all the projections into a vector $\mathbf{z}_T^\infty$ for the target and $\mathbf{z}_I^\infty$ for the interlocutor; both vectors are infinite-dimensional. Their inner product is given by a closed form,

$$\begin{aligned}\langle \mathbf{z}_T^\infty, \mathbf{z}_I^\infty \rangle &= \int_0^1 \Big(\mathbf{\Phi}(t)^T \mathbf{x}_T)^T (\mathbf{\Phi}(t)^T \mathbf{x}_I \Big) dt \\ &= \mathbf{x}_T^T \Big(\int_0^1 \mathbf{\Phi}(t) \mathbf{\Phi}(t)^T dt \Big) \mathbf{x}_I \\ &= \mathbf{x}_T^T \mathbf{G} \mathbf{x}_I,\end{aligned} \tag{1}$$

where the matrix $\mathbf{G} \in \mathcal{R}^{D \times D}$ is called the geodesic flow kernel in [2]. Since $\mathbf{G}$ is positive semidefinite, we decompose it into $\mathbf{G} = \mathbf{M}\mathbf{M}^T$ by the singular value decomposition, i.e., $\mathbf{G} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^T$ and immediately we have $\mathbf{M} = \mathbf{U}\mathbf{\Gamma}^{\frac{1}{2}}$. The new matrix $\mathbf{M} \in \mathcal{R}^{D \times D}$ is referred to as the interaction matrix [3]. As a result, instead of directly dealing with the geodesic flow, it is equivalent to transform the features using this interaction matrix. One may use the transformed features in a variety of tasks that may benefit from the mutual influence between a dyad. We will study emotion recognition in the experiments of this paper.

4.2. Weighted Geodesic Flow for Finer-Grained Interaction Modeling

In this section, we propose to improve the geodesic flow based modeling of the mutual influence effect between the interaction partners.

Motivation. Our motivation draws on Equation (1), which treats all the subspaces equally important along the geodesic flow. We argue that this could be suboptimal in the sense that the intermediate subspaces between that of the target $\mathbf{\Phi}(0)$ and that of the interlocutor $\mathbf{\Phi}(1)$ are not supported by any data at all. Instead, they are purely interpolated by the mathematical tool derived on the Grassmann manifold. However, as human communication involves a variety of dynamics and complexity over time, does the interpersonal relationship in real-life interactions fully follow such geometric imposition? Our intuition says no. The further a subspace is from $\mathbf{\Phi}(0)$ and $\mathbf{\Phi}(1)$, the less useful information it may attain of the interaction turns; in the extreme case, some of them may correspond to the noise in the data. In order to describe more detailed and expressive interaction structures, we propose a fine-grained modeling about the interactions by weighing the subspaces along the geodesic flow differently. We describe the philosophy of choosing the importance function below.

Importance Function. We propose to use an importance function $w(t), t \in [0, 1]$ to weigh the subspaces along the geodesic flow $\mathbf{\Phi}(t)$ between the dyad. It is desired to possess three properties. 1) *Positivity*: $w(t) \geq 0$ for all $t \in [0, 1]$; 2) *Normalization*: $\int_0^1 w(t)dt = 1$; and 3) *Symmetry*: $w(t) = w(1-t)$, where $t \in [0, 1]$. The symmetry property assumes both participants of a dyad equally contribute to interaction characterization. It could be removed if we had some prior knowledge that a participant is more dominant than the other in the interaction.

Immediately, the uniform importance function used in Equation 1 satisfies the above properties, so do many other functions. Mathematically, exponential and polynomial functions are more desirable since they lead to closed-form solutions to the interaction in Equation 3 with sufficient modeling flexibilities. In this work, we design the importance function $w(t)$ based on exponential functions:

$$w(t) = \frac{c}{2} \frac{e^{-ct} + e^{c(t-1)}}{1 - e^{-c}}, t \in [0, 1], \tag{2}$$

where the free parameter c controls the shape of the importance curve along the geodesic flow and characterizes detailed interaction structures between a dyad's behavior. Fig. 4 plots such importance functions with different c values. When c is far from 0, high importance is assigned to the subspaces near the two ends of the flow; when c approaches 0, $w(t)$ becomes flat so it degenerates to the case in Section 4.1 where all the subspaces have equal contributions.

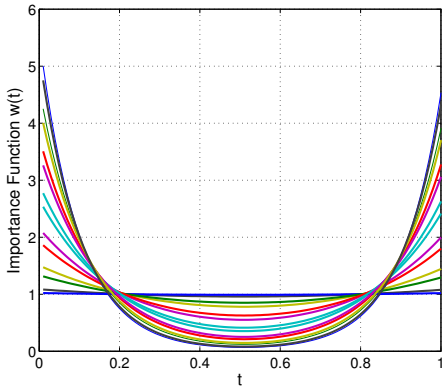


Figure 4. Importance function with different c values. Different line colors represent different c values. When c approaches 0, the importance function approaches a straight line.

Weighted Interaction Matrix. With the importance function defined, each point $\mathbf{\Phi}(t)$ on the geodesic flow is multiplied by the importance value $w(t)$ before we integrate them together. Following Equation 1, we thus arrive at a weighted geodesic flow kernel $\mathbf{G}_w$:

$$\mathbf{G}_w = \int_0^1 \big[w(t)\mathbf{\Phi}(t)\big] \big[w(t)\mathbf{\Phi}(t)\big]^T dt \tag{3}$$

Correspondingly, we further obtain the weighted interaction matrix $\mathbf{M}_w$: $\mathbf{G}_w = \mathbf{M}_w \mathbf{M}_w^T$, which describes more detailed and expressive interaction structure than the uniformly weighted one (cf. Section 4.1). In order to take advantage of the mutual influence between the dyad for emotion recognition, we transform a data point $\mathbf{x}$ by the weighted interaction matrix before sending it to a classifier.

5. Experimental Results and Analysis

We use emotion recognition as the proxy task to experimentally verify the effectiveness of the proposed weighted interaction matrix for modeling the mutual influence effect between a dyad in affective interactions. However, it is noteworthy to point out that other tasks may also benefit from our approach if they are related to the dyad's mutual influence effect. We present the main comparison results first, followed by detailed analysis about the weight function.

5.1. Emotion Recognition Results

Recall that the emotion state is manually annotated for a target turn based on multimodal behavior signals of the interaction. Our goal is to train a classifier from such data to automatically infer the emotion states from novel multimodal behaviors by investigating the mutual influence of the dyad. The classifier takes $\mathbf{f} = \mathbf{M}_w^T\mathbf{x}$ as the input, where the weighted interaction matrix $\mathbf{M}_w$ transforms the interlocutor features $\mathbf{x}$ per interaction. Here the interlocutor features represent speech (audio) and hand gesture (visual) signals of the interaction (cf. Section 3), while the new feature vector $\mathbf{f}$ induces the interaction information between a dyad. In the following experiments, we employ linear SVM as the classifier and report the average results obtained by the leave-one-interaction-out strategy.

In the interaction model, we apply principal component analysis (PCA) to identify subspaces $\mathbf{P}_T$ and $\mathbf{P}_I \in \mathcal{R}^{D\times d}$ for the target and interlocutor behavior in each interaction. The subspace dimension d is determined using cross-validation on the training set.

Baselines. For comparison, we also evaluate the same type of classifier's recognition performance using three other types of input features: 1) the plain behavioral features of the target turn (**T**), i.e., $\mathbf{x}_T$ directly extracted from the speech and gesture signals; 2) the behavioral features of both the target and interlocutor turns (**T + I**), i.e., $\mathbf{x}_T$ and $\mathbf{x}_I$, which have also been used in [24]; and 3) the geodesic flow based modeling [3] (**Geodesic**), i.e., $\mathbf{M}^T\mathbf{x}$. For 2), we follow the practice of [24] and calculate eight high level statistical functionals for each target or interlocutor turn in a dialog turn pair; they are the mean, median, standard deviation, range, lower quartile, upper quartile, minimum, and maximum from either our mapped or original behavioral features.

Results. Tables 1 and 2 respectively present the results for recognizing 2-Class and 3-Class emotions in the valence-activation space (see Fig. 3). Our approach is contrasted to the three baseline methods described above.

We draw the following observations from the tables. First of all, we find that the speech cues generally show a higher discriminative capability for distinguishing emotional dimensions than the hand gesture behavior. This is probably because the activation dimension can be better perceived from audio cues [34]. Besides, including the interlocutor information to the target turns (**T + I**) generally improves the recognition performance of target only (**T**), indicating that the interlocutor's multimodal behavior provides complementary information about the emotional state of the target subject during a dyadic interaction. These results are consistent with the findings in [24].

TABLE 1. ACCURACIES (%) FOR RECOGNIZING 2-*Class emotions* IN THE VALENCE-ACTIVATION SPACE FROM INFORMATION OF THE TARGET TURN (T), INFORMATION FROM BOTH TARGET AND INTERLOCUTOR TURNS ($T + I$) [24], AND USING INTERACTION MODELING.

Features	T	T + I	Geodesic	Ours
Audio	58.5	59.7	69.0	**70.0**
Visual	57.8	61.9	65.5	**66.4**
Audio-Visual	58.7	61.8	71.8	**73.0**

TABLE 2. ACCURACIES (%) FOR RECOGNIZING 3-*Class emotions* IN THE VALENCE-ACTIVATION SPACE FROM INFORMATION OF THE TARGET TURN (T), INFORMATION FROM BOTH TARGET AND INTERLOCUTOR TURNS ($T + I$) [24], AND USING INTERACTION MODELING.

Features	T	T + I	Geodesic	Ours
Audio	47	46	54.5	**56.5**
Visual	39.6	44	52.4	**54.0**
Audio-Visual	45.6	45	55.9	**57.0**

Finally, we note that our approach, the weighted geodesic flow based interaction model, significantly outperforms the baselines under all conditions. For example, the recognition accuracy is 58.7% and 61.8% respectively using the audio-visual information from the target turn and from both dyadic turns. The performance improves to 71.8% with the geodesic flow based and further to 73.0% with our weighted geodesic flow based model. This observation corroborates that the dyad's behavior under such modeling can serve as additional indicators of the embedded emotion in the target turn. The weighted the interaction matrix $\mathbf{M}_w$ effectively captures the dyadic behavior coordination structure in a quantitative fashion.

5.2. The Importance Function

In this section, we analyze the importance function in detail. In particular, we investigate how the shapes of the importance function $w(t)$ affect the interaction modeling and further influence the emotion recognition performance.

Fig. 5 shows the relations between the 3-Class emotion recognition performance and the free parameter c of the importance function $w(t)$ using different types of features (i.e., audio, visual, and both of them). For comparison, we include the geodesic flow based results in the figure at $c = 0$; indeed, our importance function approaches the uniform distribution between [0,1] when c goes to 0 (cf. Fig. 4).

We can see that the best results are achieved around $c = 0$ for all the three types of features. When $c < -2$ or

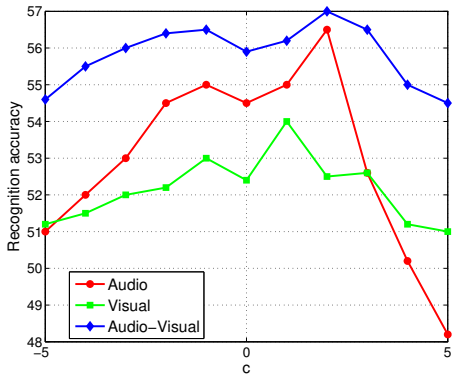


Figure 5. Accuracies (%) for recognizing 3-Class emotions using the weighted geodesic flow based model *vs.* the free parameter c in the importance function $w(t)$.

$c > 2$, the performance drops almost monotonically. This implies that the intermediate subspaces along the flow between the target and the interlocutor are actually effective in modeling their mutual influence. It also explains the inferior results of simply concatenating the features between the two (cf. **T + I** in Tables 1 and 2). Whereas the geodesic flow based modeling captures the advantages of the intermediate subspaces, our importance function provides a finer-grained way of integrating them and thus gives rise to additional gain to the emotion recognition task.

6. Conclusions and Future Work

This work focused on modeling how an interaction participant adapts his/her behavior to the multimodal behavior of the interlocutor, to express the internal emotions in a dyadic interaction. Our framework parameterized the interaction between two partners using weighted geodesic flow (WGFK) in a Grassmann manifold. Specifically, the interpersonal mutual influence is described by weighted averaging the intermediate subspaces between the behavior subspaces of an interaction dyad along the geodesic flow. Experimental results in multimodal emotion recognition tasks have shown the superiority of WGFK-based approach over the baselines in modeling expressive and detailed interaction structure.

The importance function in the WGFK-based framework offers much capacity and flexibility for modeling the complex dynamics of interaction structure. The symmetry property of the importance function (Section 4.2) can be even relaxed if the prior knowledge about which partner plays a more dominant role in an interaction is applied. In addition, adaptive learning of the importance function over time could be further incorporated in the framework as the interpersonal behavior influence evolves along the interaction.

References

[1] J. Burgoon, L. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press, 1995.

[2] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. of CVPR*, 2012.

[3] Z. Yang and S. Narayanan, "Modeling mutual influence of multimodal behavior in affective dyadic interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2234–2238.

[4] C. M. Murphy and T. J. O'Farrell, "Couple communication patterns of maritally aggressive and nonaggressive male alcoholics." *Journal of Studies on Alcohol*, vol. 58, no. 1, pp. 83–90, 1997.

[5] S. L. Johnson and T. Jacob, "Sequential interactions in the marital communication of depressed men and women." *Journal of Consulting and Clinical Psychology*, vol. 68, no. 1, p. 4, 2000.

[6] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, pp. 893–910, 1999.

[7] P. Ekman, "Body position, facial expression and verbal behavior during interviews," *Journal of Abnormal and Social Psychology*, vol. 63, 1964.

[8] R. Neumann and F. Strack, "" mood contagion": the automatic transfer of mood between persons." *Journal of personality and social psychology*, vol. 79, no. 2, p. 211, 2000.

[9] A. Kendon, "Movement coordination in social interaction: Some examples described," *Acta psychologica*, vol. 32, pp. 101–125, 1970.

[10] M. Richardson, K. Marsh, and R. Schmidt, "Effects of visual and verbal interaction on unintentional interpersonal coordination." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 1, p. 62, 2005.

[11] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proc. of ACL for Computational Linguistics: Human Language Technologies*, 2011.

[12] L. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[13] D. Heylen, E. Bevacqua, C. Pelachaud, I. Poggi, J. Gratch, and M. Schröder, "Generating listening behaviour," in *Emotion-Oriented Systems*, 2011, pp. 321–347.

[14] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Proc. of ICASSP*, 2012.

[15] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions." in *INTERSPEECH*, 2009.

[16] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," MIT Media Lab: Cambridge, MA., 2001.

[17] C.-C. Lee, A. Katsamanis, M. Black, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[18] C. Breazeal, "Regulation and entrainment in human-robot interaction," *The International Journal of Robotics Research*, vol. 21, no. 10-11, pp. 883–902, 2002.

[19] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.

[20] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples." in *INTERSPEECH*, 2010.

[21] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Proc. of ICASSP*, 2012.

[22] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 831–843, 2000.

[23] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 305–317, 2005.

[24] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, 2013.

[25] 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.

[26] "CreativeIT Database," http://sail.usc.edu/CreativeIT/.

[27] I. Guide, "Autodesk®," 2008.

[28] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," vol. 28, no. 5, p. 172, 2009.

[29] M. 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.

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

[31] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop on Speech and Emotion*, 2000.

[32] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*, 2013.

[33] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[34] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 92–105, 2011.