USE OF AFFECT BASED INTERACTION CLASSIFICATION FOR CONTINUOUS EMOTION TRACKING

Hossein Khaki, Engin Erzin

Multimedia, Vision and Graphics Lab,
Koç University, Istanbul, Turkey
hkhaki13,eerzin@ku.edu.tr

ABSTRACT

Natural and affective handshakes of two participants define the course of dyadic interaction. Affective states of the participants are expected to be correlated with the nature of the dyadic interaction. In this paper, we extract two classes of the dyadic interaction based on temporal clustering of affective states. We use the k-means temporal clustering to define the interaction classes, and utilize support vector machine based classifier to estimate the interaction class types from multimodal, speech and motion, features. Then, we investigate the continuous emotion tracking problem over the dyadic interaction classes. We use the JESTKOD database, which consists of speech and full-body motion capture data recordings of dyadic interactions with affective annotations in activation, valence and dominance (AVD) attributes. The continuous affect tracking is executed as estimation of the AVD attributes. Experimental evaluation results attain statistically significant ($p < 0.05$) improvements in affective state estimation using the interaction class information.

***Index Terms*—** Dyadic interaction type, Multimodal continuous emotion recognition, Human-computer interaction, JESTKOD database.

1. INTRODUCTION

Social signals are perceivable stimuli that, either directly or indirectly, convey information concerning social actions, interactions, attitudes, emotions and relations [1]. Through social signals of agreement and disagreement in a communicative interaction, participants can share convergent or divergent opinions, proposals, goals, attitudes and feelings. Common types of such social interactions are the group meeting scenarios [2, 3, 4, 5], political debates [6, 7, 8], theatrical improvisations [9] and broadcast conversations [10, 11].

The subject or type of a general discussion is one of the main parameters that controls the behavior of the participants in a dyadic interaction [12, 13]. Recently, multimodal dyadic interaction databases, such as CreativeIT [9] and JESTKOD [14], are available to investigate affective interaction scenarios under different interaction types. The JESTKOD database includes multimodal affective recordings of spontaneous dyadic interactions under agreement and disagreement scenarios. Moreover, this database is annotated in continuous activation-valence-dominance (AVD) space, which describes the intensity, level of pleasure, and amount of control of the emotion [15]. In our previous studies [16, 17], we investigated the role of the dyadic interaction type (DIT) on affective states and proposed a DIT based continuous emotion recognition (DIT-CER) system. In [17], the Kullback-Leibler divergence (KLD) distance of valence across agreement and disagreement interactions was found much higher than the other two attributes. Hence, we showed that valence attribute can be estimated more accurately in the DIT-CER system.

The JESTKOD database delivers a rich set of affective variability with the underlying distributions of AVD attributes. In [17], we observed the natural split of the valence distribution in agreement and disagreement interactions. Yet, one can consider to locate temporal segments, which are clustered into two classes where AVD attributes take high or low values in these binary classes, respectively. Such a general binary classification of temporal segments can provide additional information to perform CER more accurately.

In this paper, we define generalized dyadic interaction type (GIT) after binary classification of temporal windows. Then, we investigate the use of GIT for the CER problem. The reminder of this paper is organized as follows. In Section 2, we describe the proposed GIT-CER system. Then, the experimental results and discussion are given in Section 3. Finally, we conclude the paper in Section 4.

2. GIT-CER SYSTEM

Our overall GIT-CER system, which works on temporal overlapped windows is depicted in Figure 1, contains three main parts, clustering, classification, and regression. In the clustering part, each temporal window is assigned a binary class identification, as high and low, over the AVD attributes. We

This work is supported by TÜBİTAK under Grant Number 113E102.

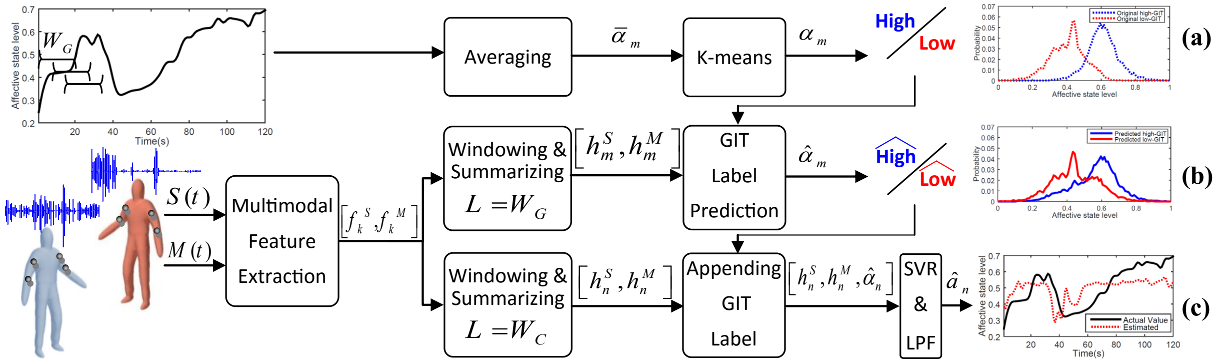


Fig. 1. Overall proposed GIT-CER system containing (a) GIT clustering, (b) GIT label prediction, and (c) CER with appended GIT label to the feature set.

refer the binary class labels as GIT label, as similarly and naturally occurs under agreement and disagreement interactions. Then, we construct estimators to predict the GIT label of a temporal window from the multimodal, speech and motion, observations in the classification part. Finally, the predicted GIT labels and the multimodal speech and motion observations are used for the CER problem in the regression part. Details of each part are described in the following.

2.1. Clustering

We target to split interaction data into two sets, in which affect attribute distributions are significantly different. Particularly, assume a_k represents one of the AVD attributes at frame k. We know that a_k varies slowly [18, 19] hence, we assign a binary class identification, as high and low, for the m'th overlapped temporal window of length W_G and rate R_G as

$$\bar{\alpha}_m = \frac{1}{W_G} \sum_{k=1+(m-1)R_G}^{W_G+(m-1)R_G} a_k, \tag{1}$$

$$\alpha_m = \mathcal{Q}(\bar{\alpha}_m), \tag{2}$$

where $\bar{\alpha}_m$ is the average of a_k over the m'th temporal window and α_m is the corresponding centroid of a two-level quantization function $\mathcal{Q}(.)$, representing the high and low affect values. We use k-means (with k=2) clustering for this binary quantizer. We refer to the quantized values as GIT labels in the reminder of this paper. Notice that these labels are predicted via multimodal features as described in Section 2.2 and we use original labels only to measure the classification error. Hence, we do not need this step directly in the test phase.

2.2. GIT Label Prediction

After dividing the data into two classes, we need to predict the label of each temporal window from speech and motion modalities. In our previous study [20], we investigate a multimodal two-class DIT estimation approach of agreement and disagreement classes from speech and motion modalities.

Same as [20], we select the acoustic frame size to have same rate as motion capture system and utilize the energy and 12-dimensional MFCC feature vector together with its first and second order derivatives for speech modality, f_k^S, and the Euler rotation angles in directions (xa,y,z) of the arm and forearm joints together with their first derivatives for motion modality, f_k^M. Then, we collect frame level feature vectors over the same overlapped temporal window as Sec 2.1 and construct matrices of feature. We fill the silence frames for the speech modality by random noise with normal distribution and construct speech feature matrix as $F_m^S = [f_1^S \cdots f_{\mathcal{M}}^S]$ for the m-th window with dimensions 39 × $\mathcal{M}$. Similarly, the motion feature matrix is constructed as $F_m^M = [f_1^M \cdots f_{\mathcal{M}}^M]$ with dimensions 24 × $\mathcal{M}$ without silence replacement.

Next, we utilize a summarization function, $F : R^{p*N} R^q$, in which p is the dimension of speech or motion features, N is the number of frames over a time window, and q is the dimension of the summarized features. Similar to [19], we utilize a variety of statistical functions such as mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, the lower and upper quantiles (corresponding to the 25th and 75th percentiles) and the inter-quantile range followed by PCA to reduce the dimension and to extract the summarized feature vector, h_m^S and h_m^M for the speech and motion modalities, respectively. Finally, we estimate the GIT label, $\hat{\alpha}_m$, for each window as

$$\hat{\alpha}_m = \Phi(h_m^S, h_m^M), \tag{3}$$

where we utilize binary linear support vector machine (SVM) as a binary predictor, i.e., $\Phi(.)$.

2.3. Continuous Emotion Recognition

CER is a regression from feature space, i.e. speech and motion, to affective state. We use same acoustic and motion features as in Section 2.2 to extract the frame-level speech, f_k^S, and motion, f_k^M, feature sets. However, since we want to estimate the exact value of a_k, we collect frame level feature vectors over shorter overlapped temporal window of length W_C and rate R_C, to construct the matrix of feature, F_n, and summarized feature vector, h_n, where n is the window index of CER. In addition, we set the mean value over the temporal window of each emotional attributes as the corresponding annotation, a_n.

Next, we append the estimated GIT label to the summarized features. Notice that since W_C is less than W_G, we interpolate the estimated GIT label $\hat{\alpha}_m$ as $\hat{\alpha}_n$ to have the same rate as the summarized features for CER. Moreover, we adjust the variance of $\hat{\alpha}_n$ equal to the first component of PCA output, to increase the contribution of GIT in regression. Finally, we apply a regression function to map the appended feature set to AVD domain as

$$\hat{a}_n = \Psi(h_n^S, h_n^M, \hat{\alpha}_n), \tag{4}$$

where $\hat{a}_n$ is the estimation of a_n and support vector regression (SVR) is used as the regression function, i.e., $\Psi(.)$.

3. EXPERIMENTAL EVALUATIONS

3.1. Exprimental Setup

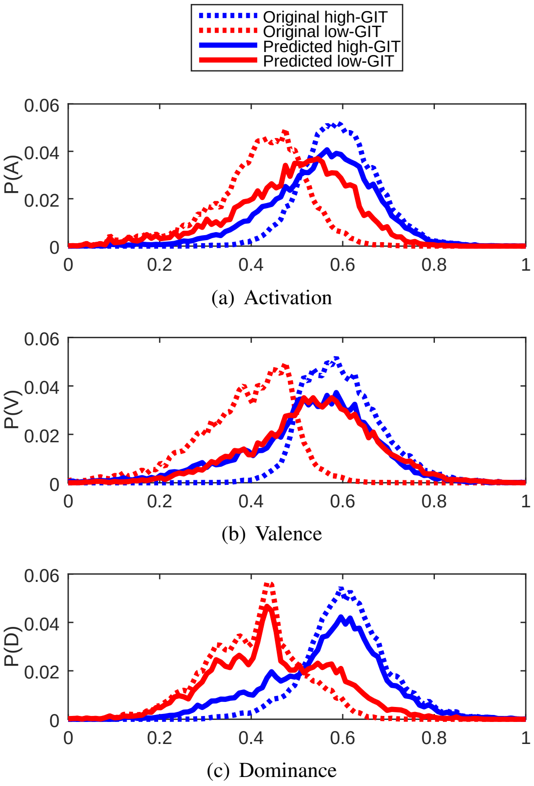
We use JESTKOD database [14] in this work, which is a multimodal database of speech, motion capture and video recordings of affective dyadic interactions. It consists of dyadic interaction recordings of 10 participants, where each participant interacted with the same partner for both agreement and disagreement settings and only appeared in one session. Annotations per attribute are performed by the same annotator. A total of six annotators contribute to collect three sets of ratings for valence and dominance, and four sets of ratings for activation attribute. Ground truth ratings are extracted from the averages over annotators' ratings.

We need to adjust the window parameters of the proposed GIT-CER system. For the CER part, we utilize same parameters as in [17], i.e., $W_C = 1.5$ s and $R_C = 0.75$ s. However, we have a trade-off for GIT parameters. Large W_G of the temporal windows leads to high quantization error and lower discrimination after clustering. On the other hand, short one results in loosing temporal properties in GIT prediction step. Hence, we perform a full search on W_G from 8 to 30 s and $R_G = 4$ s and pick the one, which achieves the highest statistical difference. To quantify the statistical difference, we utilize KLD between predicted high and low GIT distributions as

$$D_{KL}(P_H||P_L) = \sum_l P_H(l) \log \frac{P_H(l)}{P_L(l)}, \tag{5}$$

where $P_H()$ and $P_L()$ are probability distributions of each AVD over the high and low GIT labels, respectively, and l runs over the sample space of activation, valence or dominance. Moreover, we can define symmetric KLD as

$$D_{KL}(P_H, P_L) = 0.5\,(D_{KL}(P_H||P_L) + D_{KL}(P_L||P_H)).$$



(a) Activation

(b) Valence

(c) Dominance

Fig. 2. Histograms of the high and low GIT labels after clustering and multimodal prediction for each AVD attribute.

We perform the prediction and regression evaluations in a leave-one-session-out training, where we test the clips of a session at a time and train models on the remaining recordings. Hence, our test is a speaker independent evaluation. Moreover, we employ an automatic voice activity detector (VAD) [21] to replace the silent segments of the speech recordings by normal random noise and to perform evaluation over different modalities, i.e., speech (S), motion (M), and multimodal (SM), with same number of frames.

An acoustic frame is computed every 8.33 ms over 16.66 ms analysis windows to attain the same frame rate as motion capture, which is 120 fps. We adjust the PCA output dimension to preserve 90% of the total variance for the output of statistical function. We utilize binary linear support vector machine (SVM) to predict the GIT label and radial basis function kernel SVR from the LibSVM package [22] in the estimation of affective attributes.

3.2. Results and Discussion

Histograms of high and low GIT labels after clustering (original GIT) and multimodal prediction (predicted GIT) for each AVD attribute is depicted in Figure 2. Histograms of GIT classes for dominance differ significantly, they are almost separated for activation, however, valence does not convey sig-

Table 1. KLD distances between the predicted high/low GIT labels for the activation, valence and dominance attributes over speech, motion, and multimodal modalities. Values in parenthesis are selected W_G in seconds.

	$D_{KL}(P_A, P_D)$		
Modality	Activation	Valence	Dominance
Speech	0.32 (13)	0.08 (30)	0.71 (13)
Motion	0.42 (28)	0.16 (8)	0.72 (13)
Multimodal	0.31 (13)	0.11 (29)	0.81 (13)

nificant distribution differences. In addition, the symmetric KLD distances for the activation, valence and dominance attributes over different modalities are given in Table 1. Note that, we report the selected W_G in parenthesis.

Based on the Table 1, we conclude that the estimation of dominance and activation can be improved given the GIT information. However, the GIT information does not provide significant discrimination for the valence attribute. The reason can be the visual difference among the histograms, where the valence histogram of the original GIT is sharper than the others, which leads to poor prediction for the estimated GIT. Another reason can be the natural hardness of valence estimation problem from speech and motion as reported in [19].

We evaluate the performance with the correlation metric between estimated and actual affective states per each clip since variation of the AVD dimensions is more important than the exact values [19]. Finally, we calculate the mean correlation among the clips, which is depicted in Figure 3. Moreover, to illustrate how the differences between without and with predicted GIT labels are significant, we star the bars which their correlations are statistically significant ($p < 0.05$) different. In this figure, we compare the conventional CER performance with the proposed GIT-CER system. The mean correlation values of the annotators' ratings are also given as the ground truth correlation values. For activation and dominance attributes, we observe statistically significant improvements using speech, motion, and multimodal modalities. However, the difference for the other tests are not significantly different.

The yellow bars show the results of CER with given original GIT labels, which are much better than results of other tests, i.e., without or with predicted GIT labels. This difference is because of high correlation between GIT labels and AVD values, which is not available in reality. In addition, between yellow bars of each attribute, the highest bar is the one with lowest W_G, reported in Table 1. The trade-off between long and short W_G values is the main reason for this unexpected result. Since, we use original GIT labels for yellow bars, shorter W_G provides lower quantization error and higher discrimination.

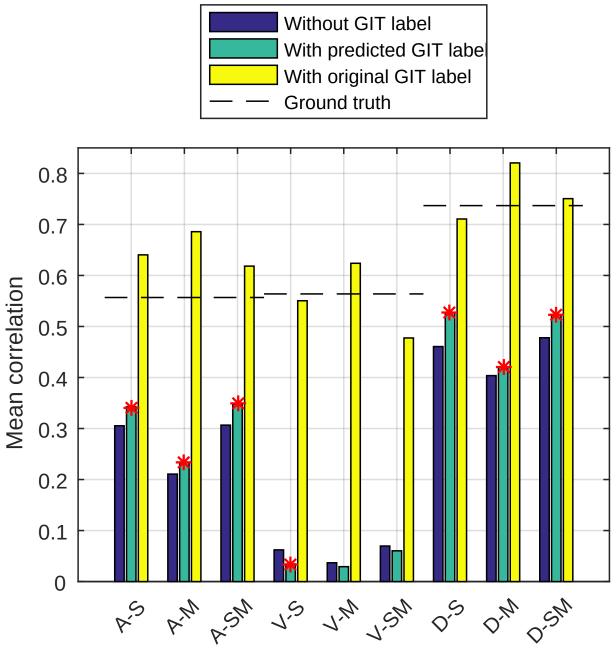


Fig. 3. Mean correlation of annotator (ground truth) and results of CER with/without original/predicted GIT labels for speech, motion, and multimodal modalities. Star signs indicate the statistically significant difference between CER without GIT labels and CER with predicted GIT labels.

4. CONCLUSION AND FUTURE WORK

In this paper, we investigate a hierarchical continuous emotion recognition system. First, we defined affect based labels, GIT labels, for the long-term temporal windows. Then, the GIT labels are predicted from speech and motion modalities. We used the predicted GIT labels together with speech and motion features to estimate the continuous emotion attributes over shorter temporal windows. The GIT labels provide useful discrimination for the activation and dominance attributes, and we significant observed improvements for the recognition of these two attributes.

Setting GIT labels to discriminate affective states better is a challenging problem and can be investigated as a future work.

5. REFERENCES

[1] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Tran. on*, vol. 3, no. 1, pp. 69–87, Jan 2012.

[2] J. Carletta, "Unleashing the killer corpus: experiences

in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[3] 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 Tran. on*, vol. 27, no. 3, pp. 305–317, March 2005.

[4] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, Stroudsburg, PA, USA, 2003, NAACL-Short '03, pp. 34–36, Association for Computational Linguistics.

[5] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004, ACL '04, Association for Computational Linguistics.

[6] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *ICASSP, 2012*, March 2012, pp. 5089–5092.

[7] K. Bousmalis, L. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Automatic Face Gesture Recognition and Workshops, 2011 IEEE International Conference on*, 2011, pp. 746–752.

[8] A. Vinciarelli, A. Dielmann, S. Favre, and H Salamin, "Canal9: a database of political debates for analysis of social interactions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009, ACII '09.

[9] A. Metallinou, Z. Yang, C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, 2015.

[10] M. Grimm, Kristian Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, June 2008, pp. 865–868.

[11] Wen Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *ICASSP, 2011*, May 2011, pp. 5556–5559.

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

[13] P. Ekman, "Body position, facial expression, and verbal behavior during interviews.," *The Journal of Abnormal and Social Psychology*, vol. 68, no. 3, pp. 295, 1964.

[14] Elif Bozkurt, Hossein Khaki, Sinan Keçeci, B. Berker Türker, Yücel Yemez, and Engin Erzin, "The jestkod database: an affective multimodal database of dyadic interactions," *Language Resources and Evaluation*, pp. 1–16, 2016.

[15] M. K Greenwald, E. W Cook, and P. J Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[16] S. Kececi, E. Erzin, and Y. Yemez, "Analysis of jestkod database using affective state annotations," in *2016 24nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2016.

[17] H. Khaki and E. Erzin, "Use of agreement/disagreement classification in dyadic interactions for continuous emotion recognition," in *INTERSPEECH*, 2016.

[18] H. Khaki and E. Erzin, "Continuous emotion tracking using total variability space," in *INTERSPEECH*, 2015.

[19] 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*, vol. 31, no. 2, pp. 137–152, 2013.

[20] H. Khaki, E. Bozkurt, and E. Erzin, "Agreement and disagreement classification of dyadic interactions using vocal and gestural cues," in *ICASSP, 2016*, March 2016, pp. 2762–2766.

[21] M. Brookes et al., "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 1997.

[22] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Tran. on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.