LIGHTLY-SUPERVISED UTTERANCE-LEVEL EMOTION IDENTIFICATION USING LATENT TOPIC MODELING OF MULTIMODAL WORDS

Zhaojun Yang and Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA

zhaojuny@usc.edu, shri@sipi.usc.edu

ABSTRACT

Research on multimodal emotion recognition has drawn much attention recently in diverse disciplines. With the increasing amount of multimodal data, unsupervised or semi-supervised learning has become highly desirable to automatically discover expression of emotion patterns in behavioral data. We present a novel approach for multimodal emotion learning using only a small amount of labels. Our approach is hinging on probabilistic latent semantic analysis (pLSA) that defines the latent variable as the emotion class, motivated by the conceptualization that human emotion acts as a latent control variable that regulates the external behavior manifestations, such as through speech and body gesture. In our approach, we represent the audio-visual information in an utterance as a bag of *multimodal words*. To exploit the interrelation between speech and gesture modalities, we propose a canonical correlation analysis (CCA) based vocabulary of multimodal words. Our approach has achieved promising experimental results. We have also demonstrated the superiority of the CCA-based multimodal words over those derived directly from the original cues.

Index Terms— Multimodal emotion recognition, unsupervised learning, semi-supervised learning, latent topic modeling

1. INTRODUCTION

The expression of emotions is inherently multimodal. It involves verbal and nonverbal behavior communicated through speech, spoken language, as well as gesture and posture of the face and body. The multimodal human behavior plays an essential role in emotion expression. Research on multimodal emotion recognition has hence received much interest recently especially with the increasing prevalence of multimodal data [1] [2]. Most of the existing approaches are based on supervised learning which requires amounts of labeled training data. However, with the increasing amount of available multimodal data [3] [4], it is tedious and expensive to obtain detailed human annotations. Moreover, the emotion annotation itself is challenging: emotion is implicitly conveyed through the external behavioral manifestations, which may result in difficulties for annotators in perceiving the hidden emotional feeling and could lead to unreliable labels. The exact emotional state however may not be needed in some applications, such as detecting emotion changes over time, where capturing relative emotion variation is desired. Unsupervised or semi-supervised learning hence is highly desirable for discovering emotion patterns from large-scale unlabeled data. The goal of this work is to develop a technique for automatically discovering the hidden emotion classes from speech and body gesture data of human interactions with only a small amount of labels.

Research efforts devoted to unsupervised or semi-supervised multimodal emotion learning thus far have been limited. Bone *et al*. proposed a knowledge-based measure for emotional arousal from prosodic features, and demonstrated its robustness across databases [5]. One weakness of such metric is the lack of generalizability in more expressive emotion varieties. Nuances of four emotion categories have been explored from postures using multivariate analysis [6]. Zhang *et al*. focused on unsupervised adaptation of acoustic features across multiple corpora for emotion recognition [7]. However, such cross-corpus adaptation requires the same type of feature or emotion inventory, which is often not available in practice.

Our work on multimodal emotion identification is inspired by the success of the unsupervised techniques in textual emotion detection [8] [9]. Among these techniques, latent topic models, which associate the latent emotion variable with the co-occurrence of words and documents, are the most popular. For example, D'Mello *et al.* applied latent semantic analysis for exploring the affect of a learner from conversational cues [8]. Promising results for textual affect recognition have been achieved by probabilistic latent semantic analysis (pLSA) in [9]. The pLSA approach introduced by Hofmann defines a proper generative model of data [10]. It allows to automatically discover latent semantic clusters from text data and to distinguish different types of word usage. This model has also been successfully applied in the challenging computer vision tasks with unlabeled images or videos, such as object detection, scene classification and human action categorization [11] [12], as well as audio classification [13].

In this work, we propose an approach hinging on pLSA for multimodal emotion learning with only a small amount of labels. The pLSA-based approach is motivated by the conceptualization that the emotion state acts as a latent control variable that regulates the external behavior manifestations, such as through speech and body gesture. In our approach, the latent variable is defined as the emotion class that governs the multimodal cues. The pLSA model is designed based on a bag of words representation. Analogous to the video representation using visual words, we transform an audio-visual utterance into a sequence of *multimodal words* and represent it as a bag of multimodal words. In order to exploit the association between modalities which jointly evolve over time during emotion expression, we develop a canonical correlation analysis (CCA) based vocabulary of multimodal words derived from the CCA transformations of speech and body gesture. We further show how to adapt the pLSA model of multimodal words learnt from unlabeled training data to a small amount of labeled data, i.e., a statistical alignment procedure to establish the cluster-emotion correspondence.

In brief, the main contribution of our work is three-fold: 1)

discovering utterance-level emotion classes in a lightly-supervised manner based on pLSA; 2) developing a CCA-based vocabulary of multimodal words for the audio-visual utterance representation; 3) presenting a statistical alignment procedure for the establishment of cluster-emotion correspondence. Our experiments show promising results, supporting that the inferred clusters from multimodal cues convey emotion information. Experimental results have also shown the superiority of the CCA-based multimodal words over those derived directly from the original cues.

2. PROPOSED APPROACH

Our goal of this work is to automatically discover utterance-level emotion classes from speech and body gesture data using only a small amount of labels. For this purpose, we propose a lightly-supervised approach hinging on pLSA that is described as below.

2.1. CCA-based Multimodal Cues

Research has shown that speech and body gesture are coherently linked to express emotions [14]. We hence exploit the association between speech and body gesture by employing canonical correlation analysis (CCA) [15] [16]. CCA is a useful statistical technique for correlating the linear relationship between two variables and has been used in a number of human-centered signal processing applications, such as the articulatory-to-acoustic mapping in speech recognition [17], bimodal fusion of facial expressions and body gesture in emotion recognition [18], and coupled information encoding of photo-sketch images in face recognition [19]. In our present problem, we have two modalities: speech $\mathbf{X} \in \mathcal{R}^{d_x \times N}$ and body gesture $\mathbf{Y} \in \mathcal{R}^{d_y \times N}$, where N is the number of samples (frames) in the dataset. CCA finds a pair of linear projections $\boldsymbol{\alpha}^T\mathbf{X}$ and $\boldsymbol{\beta}^T\mathbf{Y}$ by maximizing their correlation. Similarly, subsequent vectors $\boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_m$ can be sought by maximizing the correlation between $\boldsymbol{\alpha}_m^T\mathbf{X}$ and $\boldsymbol{\beta}_m^T\mathbf{Y}$ subject to their decorrelation with the previous ones. We hence obtain the transformation matrices for the two modalities from the CCA projection vectors, $\mathbf{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m]$ and $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m]$. A new multimodal feature set is then formed by fusing the projections, $\mathbf{F} = [\mathbf{A}^T\mathbf{X}; \mathbf{B}^T\mathbf{Y}] \in \mathcal{R}^{2m \times N}$. This new representation better captures the joint emotion information shared by modalities than the original cues $[\mathbf{X}; \mathbf{Y}]$ [17] [18].

2.2. Multimodal Codebook Creation

The latent topic model pLSA is inspired by the bag of words (BoW) representation originally proposed in the text analysis domain. BoW has also shown to be an effective and robust image or video representation in object detection and action categorization [11] [12]. It reduces data noise and transforms an entity into an effectively compact form — a histogram of words. In this work, we describe an audio-visual utterance, $u \in \{u_1, u_2, \cdots, u_U\}$, using the BoW representation as well based on a vocabulary of multimodal words (codebook) created from the multimodal data. In order to construct an effective multimodal codebook, we employ the CCA-based multimodal features $\mathbf{F}$ obtained in Section 2.1 and apply the k-means clustering method. Each resulting cluster center defines a multimodal word $w \in \{w_1, w_2, \cdots, w_V\}$ in the codebook of size V. Accordingly, each frame in an utterance is assigned to a cluster membership and is quantified as a *multimodal word*.

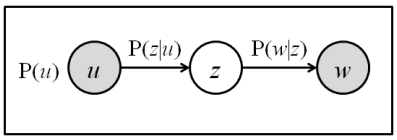


Fig. 1. The graphical model representation of pLSA [10]. The non-shaded node is the latent emotion variable z. The shaded nodes are observable variables of the utterance u and multimodal word w.

2.3. Latent Emotional Topic Model

In this section, we describe pLSA in the context of utterance-level emotion identification. pLSA is a latent variable model which associates the latent topic variable $z \in \mathcal{Z} = \{z_1, z_2, \cdots, z_K\}$ with the co-occurrence of a multimodal word w in each utterance u. The graphical model representation is illustrated in Fig. 1. Note that the topics inferred by pLSA are not restricted to the conventional semantics but dependent on the types of features. In our present problem, emotion-related multimodal features are applied (see Section 3.1), and the inferred topics hence capture affective dimensions.

Given a collection of audio-visual utterances which are described as sequences of multimodal words, the joint probability over u and w can be expressed as,

$$P(u, w) = P(u)P(w|u). \tag{1}$$

As illustrated in Fig. 1, u and w are independent conditioned on the latent variable z. Hence,

$$P(w|u) = \sum_{z \in \mathcal{Z}} P(z|u)P(w|z). \tag{2}$$

$P(z|u)$ is the probability that the emotion class z appears in a specific utterance u, and $P(w|z)$ is the probability that a multimodal word w occurs in a particular emotion class z. As seen in Eq. (2), the word distribution in a specific utterance, $P(w|u)$, is modeled as a convex combination of factors $P(w|z)$ with the mixing weights $P(z|u)$. The model parameters, $P(z|u)$ and $P(w|z)$, can be estimated by maximizing the likelihood of the multimodal words that occur in the existing utterance collection, using an expectation maximization (EM) algorithm [10].

Once we have learnt the emotion-specific word distribution, $P(w|z)$, from the unlabeled training data, we can identify the hidden emotion class given a new utterance u_{new}. According to Eq. 2, the word distribution given u_{new} is expressed as a mixture of $P(w|z)$, i.e., $P(w|u_{\text{new}}) = \sum_{z \in \mathcal{Z}} P(z|u_{\text{new}})P(w|z)$. We further estimate the mixing weights $P(z|u_{\text{new}})$ by minimizing the KL-divergence between the true distribution $P(w|u_{\text{new}})$ and the empirical one $\tilde{P}(w|u_{\text{new}})$. This procedure can also be achieved by an EM algorithm [10]. The emotion class of u_{new} is thus identified as below,

$$\hat{z}_u = \arg\max_z P(z|u_{\text{new}}). \tag{3}$$

2.4. Cluster-Emotion Correspondence

As noted in Section 2.3, each cluster z inferred from the affect-related multimodal cues corresponds to an emotion class. Since pLSA is learnt in an unsupervised manner, the emotion class that each cluster represents is still unknown. To establish the correspondence between the clusters and the emotion labels (i.e., assigning

an emotion label to each cluster), we develop a statistical alignment procedure by leveraging a small amount of labeled data.

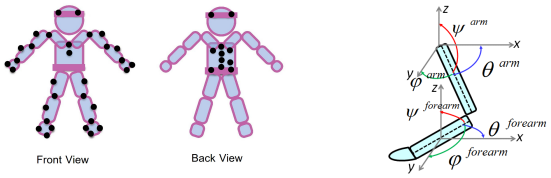
Suppose that a small subset of the training data has been manually assigned with emotion labels $e \in \{e_1, e_2, \cdots, e_K\}$. The emotion-specific empirical word distribution $P(w|e)$ can then be computed from the annotated dataset. Intuitively, the word distributions in a specific learnt cluster z and in the corresponding emotion class e, $P(w|z)$ and $P(w|e)$, are expected to shape similarly. Therefore, an optimal cluster-emotion correspondence (z, e) is uncovered when the distance between $P(w|z)$ and $P(w|e)$ is minimal. Let $\mathcal{S}$ be one permutation of the indices $\mathcal{K} = \{1, 2, \cdots, K\}$, and there are $K!$ possible permutations in total. Then, our goal is to find an optimal permutation $\hat{\mathcal{S}}$ such that the weighted average of the KL-divergence between $P(w|z_k)$ and $P(w|e_{s_k})$ is minimized,

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}} \sum_{k=1}^{K} D_{KL}(P(w|z_k)||P(w|e_{s_k})) \cdot P(e_{s_k}), \qquad (4)$$

where s_k is the k-th element in the permutation $\mathcal{S}$, and $P(e)$ is the prior of the emotion class e. The cluster-emotion consensus is henceforth determined as: $\{(z_k, e_{\hat{s}_k})|k \in \mathcal{K}\}$.

3. DATABASE DESCRIPTION

In this work, we use the USC CreativeIT database for multimodal emotion learning experiments [4]. It is a freely-available multimodal database of dyadic theatrical improvisations performed by pairs of actors. Interactions are goal-driven, which can elicit natural realization of emotions and expressive multimodal behavior. There are 50 interactions in total performed by 16 actors. The audio data of each actor was collected through close-up 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 over each actor, as shown in Fig. 2(a).



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

Fig. 2. (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 hand gesture which is the most expressive body gesture in human communication [20]. To extract hand gesture features, we manually mapped the motion data, i.e., the 3D locations of the markers, to the angles of hand joints using MotionBuilder [21]. The joint angles are popular for motion animation [22] [23] and gesture dynamics modeling [20] [24]. Fig. 2(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, energy and 12 Mel Frequency Cepstral Coefficients (MFCCs). These features were extracted every 16.67 ms (60 fps) with an analysis window length of 30 ms to match with the MoCap frame rate. The pitch features were smoothed and interpolated over the unvoiced/silence regions. We further augmented both hand gesture and acoustic features with their $1st$ derivatives. These extracted multimodal features have shown to be emotion-related and are popular in affective computing community [1] [25]. All the features were z-score normalized in a subject-dependent way.

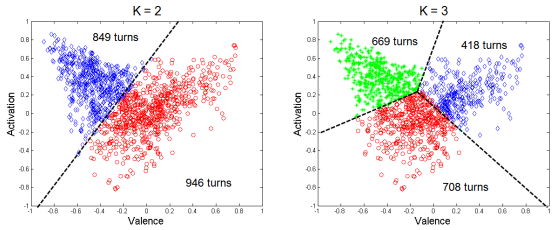


Fig. 3. The resulting emotion classes in the valence-activation space for $K = 2$ and $K = 3$.

3.2. Reference Emotion Annotation

In the database, the emotional state of each actor is annotated in terms of activation (excited *vs*. calm) and valence (positive *vs*. negative) by three or four annotators. To preserve the continuous flow of the body gesture during an interaction, time-continuous emotions for each actor are annotated throughout the recording. Annotators used Feeltrace [26] to time-continuously indicate the perceived emotion attribute value from -1 to 1 for each actor while watching the video recording (both speech and body gesture) [27].

For each actor recording, we compute the agreement (Pearson correlation) between every pair of annotators and only keep the annotator pairs with agreement greater than 0.5. Our work aims at the latent emotion discovery of utterances. Hence, we further segment each actor recording into utterances according to speech regions, resulting in 1795 utterances. The values of activation and valence of each utterance are calculated by averaging the annotations among frames within the utterance and across annotators. To provide richer and more expressive emotion varieties, we jointly consider activation and valence to create K discrete emotional classes in the valence-activation emotional space using k-means. The attribute-based emotion labels have also shown to be related to the categorical emotions [24] [28]. We consider classes with $K = 2$ and $K = 3$. Fig. 3 shows the resulting emotion classes. The emotion labels are used as ground truth for the experiment evaluation in Section 4.

4. EXPERIMENTAL RESULTS

In the experiment, we evaluate our approach by contrasting with two conventional supervised baselines in emotion recognition tasks. The first baseline represents an utterance using the popular description of utterance-level statistical functionals, such as mean, range, quantile, maximum or minimum, of the features [1]. The second baseline describes an utterance as BoW from a codebook created in the training data. Both baseline representations are used as input into a linear SVM classifier that is widely used in emotion recognition [1] [20]. In order to assess the effectiveness of the CCA-based multimodal cues, we evaluate each method respectively using only the speech features $\mathbf{X}$, the gesture features $\mathbf{Y}$, the original multimodal characteristics $[\mathbf{X}; \mathbf{Y}]$, as well as the CCA-based multimodal cues $\mathbf{F}$. Note that when using speech, gesture or the original multimodal features, the codebook is constructed in the same way as depicted in Section

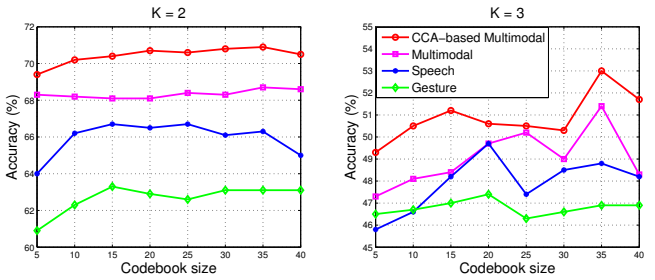


Fig. 4. Recognition accuracy using our approach with different features *vs*. codebook size V.

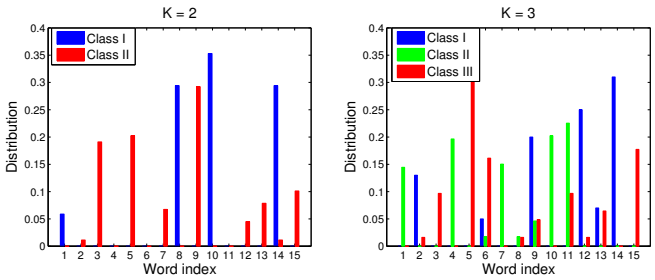


Fig. 5. The Key word distribution with codebook size 15 in each emotion class.

2.2. We adopt the leave-one-subject-out scheme for parameter selection in different methods. In each fold, the two baselines with linear SVM use all the labels in the training data, while our approach randomly keeps only 10% (an empirical value) labels in the training data for aligning cluster-emotion correspondence (see Section 2.4).

Table 1 presents the results of recognizing 2-class and 3-class emotions in the valence-activation space using different methods. Table 1 first shows that the results achieved by our approach are comparable to or even slightly better than those obtained using a linear SVM of full supervised learning. One reason could be that our approach infers the cluster structure in a non-geometrical way with respect to the co-occurrence of multimodal words. This result shows that the discovered clusters capture affective dimensions. Moreover, it is interesting to observe that the CCA-based multimodal cues outperforms the original ones with different methods. This reveals the effectiveness of CCA for exploiting the association among modalities, leading to a more informative form of multimodal description. We can also observe that the BoW descriptors improve the performance compared to the non-BoW representations under the linear SVM. Previous work has demonstrated the success of BoW in image or video representation [11]. Our results further corroborate that such an approach is also suitable for representing low-level emotion-relevant multimodal features of an utterance, especially when combining with the CCA-based multimodal codebook. Another observation from Table 1 is that the multimodal cues are more informative about emotions in contrast to the speech or gesture only information.

We investigate the effect of the codebook size on the recognition performance of our approach. Fig. 4 shows the relation of the recognition performance and the codebook size using different types of cues. When using the CCA-based multimodal words, the best performance in both 2-class and 3-class recognition tasks is achieved with the codebook size 35. We also notice that a more compact codebook is required for the single modality of speech or gesture to perform well, compared to the multimodal case.

Table 1. Accuracies (%) for recognizing 2-Class and 3-Class emotions in the valence-activation space using different methods.

Method	Feature	$K=2$	$K=3$
Non-BoW (SVM)	Gesture	61.8	42.1
	Speech	62.7	46.5
	Multimodal	63.6	48.0
	CCA-based Multimodal	**64.6**	**49.3**
BoW (SVM)	Gesture	62.5	45.6
	Speech	66.1	49.8
	Multimodal	66.6	50.8
	CCA-based Multimodal	**68.1**	**51.0**
Our approach	Gesture	63.3	47.4
	Speech	66.0	49.7
	Multimodal	68.7	51.4
	CCA-based Multimodal	**70.9**	**53.0**

5. ANALYSIS OF MULTIMODAL WORD USAGE

One auxiliary benefit of the pLSA model is that it provides the emotion-specific word distribution $P(w|z)$ and utterance distribution $P(u|z)$ which could be used to localize the emotion-dependent key words (frames) within an utterance. Analysis of the identified key words with respect to a specific emotion is helpful for understanding the multimodal characteristics in each emotion category. Similarly to identifying the utterance-level emotion class, we could assign an emotion label to each word within the utterance according to the posteriors $P(z|w,u)$,

$$\hat{z}_w = \arg\max_z P(z|w,u),\ P(z|w,u) = \frac{P(w|z)P(z|u)}{\sum_z P(w|z)P(z|u)}. \quad (5)$$

The key words per utterance are thus localized when they are assigned to the same utterance-level emotion class.

As an example, we compute the histogram of identified key words in each emotion category using the CCA-based multimodal codebook of size 15. The results are presented in Fig. 5. Bars of different colors represent key word distributions in different emotion classes. A clear observation is the discrimination of the distributions of key words in distinct emotion classes. Let's take the 2-class emotions in Fig. 5(a) for instance. The popular key words in class I are 8, 10 and 14, while 3, 5 and 9 are dominant in class II. Some key words are also shared among distinct classes, such as 6 and 9 in Fig. 5(b). This analysis sheds light into the understanding of the interplay between multimodal behavior and the emotion class. It also explains the effectiveness of our approach for emotion identification as demonstrated in the results of Table 1.

6. CONCLUSION AND FUTURE WORK

In this work, we presented an approach hinging on pLSA for multimodal emotion learning with light supervision using a small amount of labeled data. This approach can be readily applied for emotion change detection over time, a topic for future work. Also, in the future, we plan to extend this model by incorporating the cluster-emotion alignment procedure in the learning stage. We could also consider using advanced nonlinear mapping techniques, such as deep neural network [29], instead of CCA for multimodal modeling.

7. REFERENCES

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C-M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. of international conference on Multimodal interfaces*, 2004, pp. 205–211.

[2] 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.

[3] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J-C. Martin, L. Devillers, S. Abrilian, and A. Batliner, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*, pp. 488–500. Springer, 2007.

[4] A. Metallinou, Z. Yang, C-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*, pp. 1–25, 2015.

[5] D. Bone, C. Lee, and S. S. Narayanan, "A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation.," in *Proc. of INTERSPEECH*, 2012.

[6] P R. De Silva, A. Kleinsmith, and N. Bianchi-Berthouze, "Towards unsupervised detection of affective body posture nuances," in *Affective Computing and Intelligent Interaction*, pp. 32–39. Springer, 2005.

[7] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Automatic Speech Recognition and Understanding*. IEEE, 2011, pp. 523–528.

[8] S. K. Dmello, S. D. Craig, A. Witherspoon, B. Mcdaniel, and A. Graesser, "Automatic detection of learners affect from conversational cues," *User modeling and user-adapted interaction*, vol. 18, no. 1-2, pp. 45–80, 2008.

[9] S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," in *Proc. of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 62–70.

[10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[11] J. C. Niebles, H. Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[12] J. Sivic, B. C. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. of ICCV*, 2005, vol. 1, pp. 370–377.

[13] S. Kim, PG. Georgiou, and S. Narayanan, "Latent acoustic topic models for unstructured audio classification," *APSIPA Transactions on Signal and Information Processing*, vol. 1, no. 6, pp. 1–15, 2012.

[14] Z. Yang and S. Narayanan, "Analysis of emotional effect on speech-body gesture interplay," in *Proc. of Interspeech*, 2014.

[15] P K. Atrey, M A. Hossain, A. El Saddik, and M S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[16] B. Thompson, "Canonical correlation analysis," *Encyclopedia of statistics in behavioral science*, 2005.

[17] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements.," in *MLSLP*, 2012, pp. 34–37.

[18] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures.," in *Proc. of BMVC*, 2007, pp. 1–10.

[19] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. of CVPR*. IEEE, 2011, pp. 513–520.

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

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

[22] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 172, 2009.

[23] 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.

[24] Z. Yang and S. Narayanan, "Modeling mutual influence of multimodal behavior in affective dyadic interactions," in *Proc. of ICASSP*, 2015, pp. 2234–2238.

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

[26] 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.

[27] 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.

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

[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. of ICML*, 2011, pp. 689–696.