Multimodal Human Behavior Analysis: Learning Correlation and Interaction Across Modalities

Yale Song
MIT CSAIL
Cambridge, MA 02139
yalesong@csail.mit.edu

Louis-Philippe Morency
USC ICT
Los Angeles, CA 90094
morency@ict.usc.edu

Randall Davis
MIT CSAIL
Cambridge, MA 02139
davis@csail.mit.edu

ABSTRACT

Multimodal human behavior analysis is a challenging task due to the presence of complex nonlinear correlations and interactions across modalities. We present a novel approach to this problem based on Kernel Canonical Correlation Analysis (KCCA) and Multi-view Hidden Conditional Random Fields (MV-HCRF). Our approach uses a nonlinear kernel to map multimodal data to a high-dimensional feature space and finds a new projection of the data that maximizes the correlation across modalities. We use a multi-chain structured graphical model with disjoint sets of latent variables, one set per modality, to jointly learn both view-shared and view-specific substructures of the projected data, capturing interaction across modalities explicitly. We evaluate our approach on a task of agreement and disagreement recognition from nonverbal audio-visual cues using the Canal 9 dataset. Experimental results show that KCCA makes capturing nonlinear hidden dynamics easier and MV-HCRF helps learning interaction across modalities.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*

Keywords

Multimodal signal processing; multi-view latent variable discriminative models; canonical correlation analysis; kernel methods

1. INTRODUCTION

Human communication is often accompanied by multimodal nonverbal cues, such as gestures, eye gaze, and facial expressions. These nonverbal cues play an important role in the way we communicate with others and can convey as much information as spoken language. They complement or substitute for spoken language, help to illustrate or emphasize main points, and provide a rich source of predictive information for understanding the intentions of the others. Automatic analysis of human behavior can thus benefit from harnessing multiple modalities.

From the machine learning point of view, multimodal human behavior analysis continues to be a challenging task, in part because learning the complex relationship across modalities is non-trivial. Figure 1(a) shows a pair of time-aligned sequences with audio and visual features (from [11]; details can be found in Section 4.1). When learning with this type of data, it is important to consider the correlation and interaction across modalities: An underlying correlation structure between modalities may make the amplitude of the signal from one modality different in relation to the signal from another modality, e.g., loud voice with exaggerated gestures. Also, the interaction between modalities may have certain patterns that change the direction in which each sequence may evolve over time.

In this paper, we investigate the hypothesis that transforming the original data to be maximally correlated across modalities, in a statistical sense, and capturing the interaction across modalities explicitly from the transformed data improves recognition performance on human behavior analysis. To this end, we present a novel approach to multimodal data learning that captures complex nonlinear correlations and interactions across modalities, based on KCCA [3] and MV-HCRF [10]. Our approach uses a nonlinear kernel to map multimodal data to a high-dimensional feature space and finds a new projection of the data that maximizes the correlation across modalities. Figure 1(b) shows the projected signals found by KCCA, where the relative importance of gestures 'head shake' and 'shoulder shrug' have been emphasized to make the statistical relevance between the audio and visual signals become as clear as possible. We then capture the interaction across modalities using a multi-chain structured latent variable discriminative model. The model uses disjoint sets of latent variables, one set per view, and jointly learns both view-shared and view-specific sub-structures of the projected data.

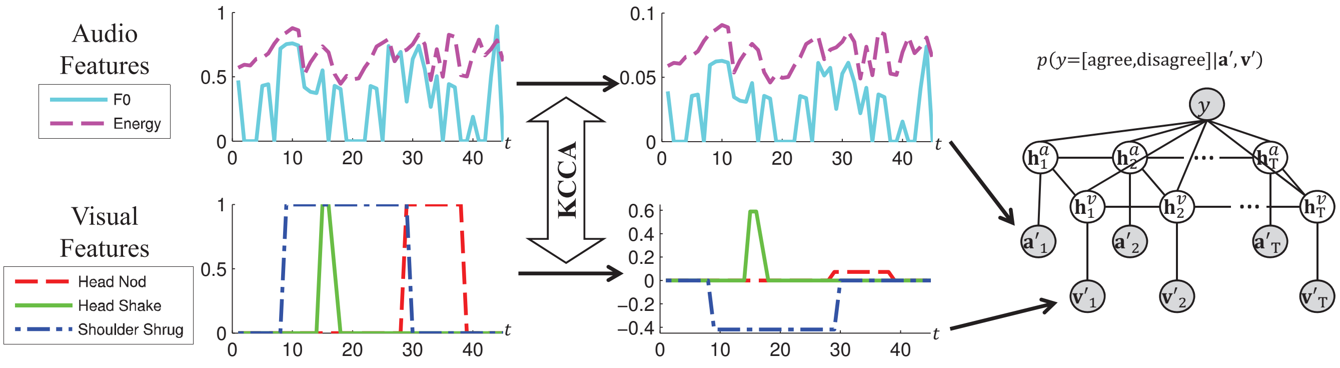
We evaluated our approach using the Canal 9 dataset [11], where the task is to recognize agreement and disagreement from nonverbal audio-visual cues. We report that using KCCA with MV-HCRF to learn correlation and interaction across modalities successfully improves recognition performance compared to baseline methods.

2. RELATED WORK

Due to its theoretical and practical importance, multimodal human behavior analysis has been a popular research topic. While audio-visual speech recognition is probably the most well known and successful example [6], multimodal affect recognition has recently been getting considerable attention [12]. Bousmalis *et al.* [1] proposed a system for spontaneous agreement and disagreement recognition based only on prosodic and gesture cues, as we did here. They used an HCRF to capture the hidden dynamics of the multimodal cues. However, their approach did not consider the correlation and interaction across modalities explicitly.

Canonical correlation analysis (CCA) has been successfully applied to multimedia content analysis [3, 13]. Hardoon *et al.* [3]

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.



(a) Original Data (b) Projected Data (c) MV-HCRF

Figure 1: An overview of our approach. (a) An audio-visual observation sequence from the Canal 9 dataset [11]. KCCA uses a nonlinear kernel to map the original data to a high-dimensional feature space, and finds a new projection of the data in the feature space that maximizes the correlation between audio and visual channels. (b) The projected data shows that emphasizing the amplitude of the 'head shake' and 'shoulder shrug' gestures maximized the correlation between audio and visual channels. (c) multi-view HCRF for jointly learning both view-shared and view-specific sub-structures of the projected data. $\mathbf{a}_t$ and $\mathbf{v}_t$ are observation variables from audio and visual channels, and h^a_t and h^v_t are hidden variables for audio and visual channels.

used kernel CCA (KCCA) for learning the semantic representation of images and their associated text. However, their approach did not consider capturing hidden dynamics in the data. Latent variable discriminative models, e.g., HCRF [7], have shown promising results in human behavior analysis tasks, for their ability to capture the hidden dynamics (e.g., spatio-temporal dynamics). Recently, the multi-view counterpart [10] showed a significant improvement over single-view methods in recognizing human actions. However, their work did not learn nonlinear correlation across modalities. We extend this body of work, enabling it to modeling multimodal human behavior analysis.

3. OUR APPROACH

Consider a labeled sequential dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}_i$ is a multivariate observation sequence and $y_i \in \mathcal{Y}$ is a sequence label from a finite label set $\mathcal{Y}$. Since we have audio-visual data, we use the notation $\mathbf{x}_i = (\mathbf{a}_i, \mathbf{v}_i)$ where $\mathbf{a}_i \in \mathbb{R}^{n_a \times T}$ and $\mathbf{v}_i \in \mathbb{R}^{n_v \times T}$ are audio and visual sequences of length T, respectively.

Figure 1 shows an overview of our approach. We first find a new projection of the original observation sequence $\mathbf{x}' = (\mathbf{a}', \mathbf{v}')$ using KCCA [3] such that the correlation between $\mathbf{a}'$ and $\mathbf{v}'$ is maximized in the projected space (Section 3.1). Then we use this projected data as an input to MV-HCRF [10] to capture hidden dynamics and interaction between audio and visual data (Section 3.2).

3.1 KCCA

Given a set of paired samples $\{(\mathbf{a}_i, \mathbf{v}_i)\}_{i=1}^N$, $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$ and $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]$, Canonical Correlation Analysis (CCA) aims to find a pair of transformations $\mathbf{w}_a$ and $\mathbf{w}_v$ such that the correlation between the corresponding projections $\rho(\mathbf{w}_a^\top \mathbf{A}, \mathbf{w}_v^\top \mathbf{V})$ is maximized. However, since CCA finds $\mathbf{w}_a$ and $\mathbf{w}_v$ that are linear in the vector space, it may not reveal nonlinear relationships in the data [9].

Kernel CCA (KCCA) [3] uses the kernel trick [9] to overcome this limitation by projecting the original data onto a high dimensional feature space before running CCA. A kernel is a function $K(\mathbf{x}_i, \mathbf{x}_j)$ that, for all $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}$,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product, and Φ is a nonlinear mapping function to a Hilbert space $\mathcal{F}$, $\Phi : \mathbf{x} \in \mathbb{R} \mapsto \Phi(\mathbf{x}) \in \mathcal{F}$.

To apply the kernel trick, the standard KCCA then rewrites $\mathbf{w}_a$ (and $\mathbf{w}_v$) as an inner product of the data $\mathbf{A}$ (and $\mathbf{V}$) with a direction α (and β),

$$\mathbf{w}_a = \mathbf{A}^\top \alpha, \quad \mathbf{w}_v = \mathbf{V}^\top \beta \tag{1}$$

If we assume that $\mathbf{a}$'s and $\mathbf{v}$'s are centered (i.e., mean zero), the goal is to maximize the correlation coefficient

$$\begin{aligned}
\rho(\cdot,\cdot) &= \max_{\mathbf{w}_a, \mathbf{w}_v} \frac{\mathbb{E}[\mathbf{w}_a^\top \mathbf{a}\mathbf{v}^\top \mathbf{w}_v]}{\sqrt{\mathbb{E}[\mathbf{w}_a^\top \mathbf{a}\mathbf{a}^\top \mathbf{w}_a]}\sqrt{\mathbb{E}[\mathbf{w}_v^\top \mathbf{v}\mathbf{v}^\top \mathbf{w}_v]}} \\
&= \max_{\mathbf{w}_a, \mathbf{w}_v} \frac{\mathbf{w}_a^\top \mathbf{A}\mathbf{V}^\top \mathbf{w}_v}{\sqrt{\mathbf{w}_a^\top \mathbf{A}\mathbf{A}^\top \mathbf{w}_a \mathbf{w}_v^\top \mathbf{V}\mathbf{V}^\top \mathbf{w}_v}} \\
&= \max_{\alpha, \beta} \frac{\alpha \mathbf{A}\mathbf{A}^\top \mathbf{V}\mathbf{V}^\top \beta}{\sqrt{\alpha \mathbf{A}\mathbf{A}^\top \mathbf{A}\mathbf{A}^\top \alpha \cdot \beta \mathbf{V}\mathbf{V}^\top \mathbf{V}\mathbf{V}^\top \beta}} \\
&= \max_{\alpha, \beta} \frac{\alpha^\top K_a K_v \beta}{\sqrt{\alpha^\top K_a^2 \alpha \cdot \beta^\top K_v^2 \beta}}.
\end{aligned} \tag{2}$$

where $K_a = K(\mathbf{A}, \mathbf{A})$ and $K_v = K(\mathbf{V}, \mathbf{V})$ are kernel matrices.

Since Equation 2 is scale invariant with respect to α and β (they cancel out), the optimization problem is equivalent to:

$$\max_{\alpha, \beta} \; \alpha^\top K_a K_v \beta \quad \text{subject to} \quad \alpha^\top K_a^2 \alpha = \beta^\top K_v^2 \beta = 1 \tag{3}$$

The corresponding Lagrangian dual form is

$$L(\alpha, \beta, \theta) = \alpha^\top K_a K_v \beta - \frac{\theta_\alpha}{2}(\alpha^\top K_a^2 \alpha - 1) - \frac{\theta_\beta}{2}(\beta^\top K_v^2 \beta - 1) \tag{4}$$

The solution to Equation 3 is found by taking derivatives of Equation 4 with respect to α and β, and solving a standard eigenvalue problem [5]. However, when K_a and K_v are non-invertible, as is common in practice with large datasets, problems can arise such as computational issues or degeneracy. This problem is solved by applying the partial Gram-Schmidt orthogonalization (PGSO) with a precision parameter η to reduce the dimensionality of the kernel matrices and approximate the correlation.

After we find α and β, we plug the solution back in to Equation 1 to obtain $\mathbf{w}_a$ and $\mathbf{w}_v$, and finally obtain new projections:

$$\begin{aligned}\mathbf{A}' &= [\langle \mathbf{w}_a, \mathbf{a}_1\rangle, \cdots, \langle \mathbf{w}_a, \mathbf{a}_N\rangle] \\ \mathbf{V}' &= [\langle \mathbf{w}_v, \mathbf{v}_1\rangle, \cdots, \langle \mathbf{w}_v, \mathbf{v}_N\rangle].\end{aligned} \quad (5)$$

3.2 Multi-view HCRF

Given the new projection's audio-visual features $\mathbf{A}'$ and $\mathbf{V}'$ (Equation 5), the next step is to learn the hidden dynamics and interaction across modalities (see Figure 1 (b) and (c)).

A Multi-View Hidden Conditional Random Field [10] (MV-HCRF) is a conditional probability distribution that factorizes according to a multi-chain structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each chain is a discrete representation of each view. We use disjoint sets of latent variables $\mathbf{h}^a \in \mathcal{H}^a$ for audio and $\mathbf{h}^v \in \mathcal{H}^v$ for visual channel to learn both view-shared and view-specific sub-structures in audio-visual data. An MV-HCRF defines $p(y \mid \mathbf{a}', \mathbf{v}')$ as

$$p(y \mid \mathbf{a}', \mathbf{v}') = \frac{\sum_{\mathbf{h}} \exp\{\Lambda^{\intercal}\Phi(y, \mathbf{h}, \mathbf{a}', \mathbf{v}')\}}{\sum_{y', \mathbf{h}} \exp\{\Lambda^{\intercal}\Phi(y', \mathbf{h}, \mathbf{a}', \mathbf{v}')\}}$$

where $\mathbf{h} = \{\mathbf{h}^a, \mathbf{h}^v\}$ and $\Lambda = [\lambda, \omega]$ is a model parameter vector. The function $\Lambda^{\intercal}\Phi(y, \mathbf{h}, \mathbf{a}', \mathbf{v}')$ is factorized with feature functions $f_k(\cdot)$ and $g_k(\cdot)$ as

$$\begin{aligned}\Lambda^{\intercal}\Phi(y, \mathbf{h}, \mathbf{a}', \mathbf{v}') &= \sum_{s \in \mathcal{V}} \sum_k \lambda_k f_k(y, h_s, \mathbf{a}', \mathbf{v}') \\ &+ \sum_{(s,t) \in \mathcal{E}} \sum_k \omega_k g_k(y, h_s, h_t, \mathbf{a}', \mathbf{v}').\end{aligned}$$

Following [10], we define three types of feature functions. The *label* feature function $f_k(y, h_s)$ encodes the relationship between a latent variable h_s and a label y. The *observation* feature function $f_k(h_s, \mathbf{a}', \mathbf{v}')$ encodes the relationship between a latent variable h_s and observations $\mathbf{x}$. The *edge* feature function $g_k(y, h_s, h_t)$ encodes the transition between two latent variables h_s and h_t. We use the linked topology from [10] to define the edge set $\mathcal{E}$ (shown in Figure 1(c)), which models contemporaneous connections between audio and visual observations, i.e., the concurrent latent states in the audio and visual channel mutually affect each other. Note that the $f_k(\cdot)$ are modeled under the assumption that views are conditionally independent given respective sets of latent variables, and thus encode the view-specific sub-structures. The feature function $g_k(\cdot)$ encodes both view-shared and view-specific sub-structures.

The optimal parameter set Λ^* is found by minimizing the negative conditional log-likelihood

$$\min_{\Lambda} L(\Lambda) = \frac{1}{2\sigma^2} \|\Lambda\|^2 - \sum_{i=1}^{N} \log p(y_i \mid \mathbf{a}'_i, \mathbf{v}'_i;\ \Lambda) \quad (6)$$

where the first term is the Gaussian prior over Λ that works as an L_2-norm regularization. We find the optimal parameters Λ^* using gradient descent with a quasi-newton optimization method, the limited-memory BFGS algorithm [5]. The marginal probability of each node is obtained by performing an inference task using the Junction Tree algorithm [2].

4. EXPERIMENT

In this section, we describe the dataset, detail our experimental methodology, and discuss our results.

4.1 Dataset

We evaluated our approach using the Canal9 dataset [11], where the task is to recognize agreement and disagreement from nonverbal audio-visual cues during spontaneous political debates. The Canal9 dataset is a collection of 72 political debates recorded by the Canal 9 TV station in Switzerland, with a total of roughly 42 hours of recordings. In each debate there is a moderator and two groups of participants who argue about a controversial political question. The dataset contains highly spontaneous verbal and non-verbal multimodal human behavior data.

In order to facilitate comparison, we used the same part of the Canal9 dataset with nonverbal audio-visual features as was selected for use in Bousmalis *et al.* [1]. This consisted of 53 episodes of agreements and 94 episodes of disagreement collected over 11 debates. Bousmalis *et al.* selected the episodes based on two criteria: (a) the verbal content clearly indicates agreement/disagreement, which ensures that the ground truth label for the episode is known; (b) the episode includes only one person, with a close-up shot of the speaker. The audio channel was encoded with 2-dimensional prosodic features, including the fundamental frequency (F0) and the energy. The visual channel was encoded with 8 gestures: *'Head Nod', 'Head Shake', 'Forefinger Raise', 'Forefinger Raise-Like', 'Forefinger Wag', 'Hand Wag', 'Hands Scissor', and 'Shoulder Shrug'*, where the presence/absence of the actions in each frame was manually annotated with binary values. We downsampled the data from the original sampling rate of 25 *kHz* to 12.5 *kHz*.

4.2 Methodology

The first step in our approach is to run KCCA to obtain a new projection of the data. We used a Gaussian RBF kernel as our kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\| / 2\gamma^2\right)$ because of its empirical success in the literature [9]. We validated the kernel width $\gamma = 10^k, k = [-1, 0, 1]$ and the PGSO precision parameter $\eta = [1 : 6]$ using grid search. The optimal parameter values were chosen based on the maximum correlation coefficient value.

Our experiments followed a leave-two-debates-out cross-validation approach, where we selected 2 debates of the 11 debates as the test split, 3 debates for the validation split, and the remaining 6 debates for the training split. This was repeated five times on the 11 debates. The F1 scores were averaged to get the final result. We chose four baseline models: Hidden Markov Models (HMM) [8], Conditional Random Fields (CRF) [4], Hidden Conditional Random Fields (HCRF) [7], and Multi-view HCRF (MV-HCRF) [10]. We compared this to our KCCA with MV-HCRF approach. Note that HMM and CRF perform per-frame classification, while HCRF and MV-HCRF perform per-sequence classification. The classification results of each model in turn were measured accordingly.

We automatically validated the hyper-parameters of all models. For all CRF-family models, we varied the L_2-norm regularization factor $\sigma = 10^k, k = [0, 1, 2]$ (see Equation 6). For HMM and HCRF, the number of hidden states were varied $|\mathcal{H}| = [2 : 6]$; for MV-HCRF, they were $\left(|\mathcal{H}^A|, |\mathcal{H}^V|\right) = ([2 : 4], [2 : 4])$. Since the optimization problems in HMM, HCRF and MV-HCRF are non-convex, we performed two random initializations of each model; the best model was selected based on the F1 score on the validation split. The L-BFGS solver was set to terminate after 500 iterations.

4.3 Result and Discussion

We first compared our approach to existing methods: HMM [8], CRF [4], HCRF [7], and MV-HCRF [10]. Figure 2 shows a bar plot of mean F1 scores and their standard deviations. We can see that our approach clearly outperforms all other models.

For further analysis, we investigated whether learning nonlinear correlation was important, comparing KCCA to CCA and the original data. Table 1 shows that models trained with KCCA always outperformed the others, suggesting that learning nonlinear correlation in the data was important. Figure 1(b) shows the data projected in

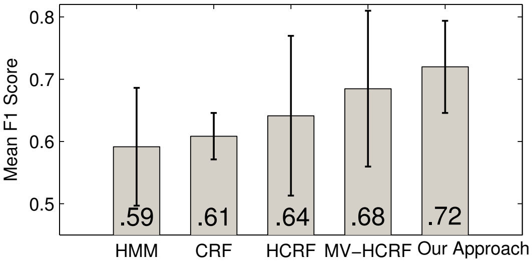


Figure 2: A bar plot of mean F1 scores with error bars showing standard deviations. This shows empirically that our approach successfully learned correlations and interactions between audio and visual features using KCCA and MV-HCRF.

Models	Original Data	CCA	KCCA
HMM	.59 (.09)	.59 (.12)	.61 (.13)
CRF	.61 (.04)	.63 (.03)	.67 (.08)
HCRF	.64 (.13)	.65 (.07)	.69 (.06)
MV-HCRF	**.68 (.13)**	**.71 (.07)**	**.72 (.07)**

Table 1: Experimental results (means and standard deviations of F1 scores) comparing KCCA to CCA and the original data. The results show that learning nonlinear correlation in the data was important in our task.

a new space found by KCCA, where the ‘head shake’ and ‘shoulder shrug’ gestures were relatively emphasized compared to ‘head nod’, which maximized the correlation between the audio and visual signals. We believe that this made our data more descriptive, allowing the learning algorithm to capture the hidden dynamics and interactions between modalities more effectively.

We also investigated whether our approach captures interaction between audio-visual signals successfully. We compared the models trained with a unimodal feature (audio or visual) to the models trained with audio-visual features. Table 2 shows means and standard deviations of the F1 scores. In the three single-chain models, HMM, CRF, and HCRF, there was an improvement when both audio and visual features were used, confirming that using a combination of audio and visual features for our task is indeed important. Also, MV-HCRF outperformed HMM, CRF, HCRF, showing empirically that learning interaction between audio-visual signals explicitly improved the performance.

5. CONCLUSIONS

We presented a novel approach to multimodal human behavior analysis using KCCA and MV-HCRF, and evaluated it on a task of recognizing agreement and disagreement from nonverbal audio-visual cues using the Canal9 dataset. On this dataset, we showed that preprocessing multimodal data with KCCA, by projecting the data to a new space where the correlation across modalities is maximized, helps capture complex nonlinear relationship in the data. We also showed that KCCA with MV-HCRF, which jointly learns view-shared and view-specific interactions explicitly, improves the recognition performance, showing that our approach successfully captured complex nonlinear interaction across modalities. We look forward to applying our technique to other applications in multimodal human behavior analysis for further analysis.

Models	Audio	Video	Audio+Video
HMM	.54 (.08)	.58 (.11)	**.59 (.09)**
CRF	.48 (.05)	.58 (.15)	**.61 (.04)**
HCRF	.52 (.09)	.60 (.09)	**.64 (.13)**
MV-HCRF	·	·	**.68 (.13)**
KCCA + MV-HCRF	·	·	**.72 (.07)**

Table 2: Experimental results (means and standard deviations of F1 scores) comparing unimodal (audio or video) features to the audio-visual features. The results confirms that using both audio and visual features are important in our task.

6. ACKNOWLEDGMENTS

This work was funded by ONR grant #N000140910625, NSF grant #IIS-1118018, NSF grant #IIS-1018055, and U.S. Army RDECOM.

7. REFERENCES

[1] K. Bousmalis, L.-P. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *FG*, 2011.

[2] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

[3] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comp.*, 16(12):2639–2664, 2004.

[4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[5] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

[6] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. *Audio-Visual Automatic Speech Recognition: An Overview*. MIT Press, 2004.

[7] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *TPAMI*, 29(10):1848–1852, 2007.

[8] L. R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., 1990.

[9] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[10] Y. Song, L.-P. Morency, and R. Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*, 2012.

[11] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *ACII*, 2009.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *TPAMI*, 31(1):39–58, 2009.

[13] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *MM*, pages 273–276, 2007.