Beyond Facial Expressions: Learning Human Emotion from Body Gestures

Caifeng Shan, Shaogang Gong, and Peter W. McOwan
Department of Computer Science
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
{cfshan, sgg, pmco}@dcs.qmul.ac.uk

Abstract

Vision-based human affect analysis is an interesting and challenging problem, impacting important applications in many areas. In this paper, beyond facial expressions, we investigate affective body gesture analysis in video sequences, a relatively understudied problem. Spatial-temporal features are exploited for modeling of body gestures. Moreover, we present to fuse facial expression and body gesture at the feature level using Canonical Correlation Analysis (CCA). By establishing the relationship between the two modalities, CCA derives a semantic "affect" space. Experimental results demonstrate the effectiveness of our approaches.

1 Introduction

The ability to recognize affective states of a person is indispensable and important for successful interpersonal social interaction. Affective arousal modulates all nonverbal communication cues such as facial expression, body moment and posture, gesture, and tone of voice. Design and development of an automated system that can detect and interpret human affective behavior is an interesting and challenging problem [22], impacting important applications in many areas.

In computer vision, affect analysis from facial expression has been widely studied in recent years [23, 8]. However, little attention has been placed on affective body posture and gesture analysis (see Figure 1 for examples of affective body gesture), although bodily expression plays a vital role in conveying human emotional states, and the perception of facial expression is strongly influenced by the concurrently presented body language [1, 19]. This is probably due to the high variability of the emotional body posture and gesture that can be displayed. The existing studies on vision-based gesture recognition have been primarily carried out on non-affective gestures such as sign languages [27]. Emotional bodily expression has been studied in psychology and non-verbal communication. For example, Coulson [5] presented experiments on attributing six universal emotions to static body postures using computer-generated mannequin figures, and his experiments suggest that recognition from body posture is comparable to recognition from the voice, and some postures are recognized as well as facial expressions. Statistical techniques were used in [25] to determine a set of posture features in discriminating between emotions. Burgoon *et al.*[4] discussed the issue of identifying emotional states from bodily

cues for human behavior understanding. Mota and Picard [15] studied affective postures in an e-learning scenario, where the posture information was collected through a sensor chair. An affective gesture recognition system has been introduced in [26] to recognize children's emotion with intensity in context of game. Recently some tentative attempts have been made on vision-based affective body gesture analysis [2, 11]. However, there are some limitations in these studies. For example, they used very limited data (for instance, only 27 video sequences from 4 subjects were processed in [11]), and the feature extraction and representation are rather simple (for instance, the neutral and expressive frames are manually selected in [11]).

Figure 1: Examples of affective body gestures (from the FABO database [10]). From *top* to *bottom*: Fear, Joy, Uncertainty, and Surprise.

The face and the body, as part of an integrated whole, both contribute in conveying the emotional state of the individual. The studies in psychology [1, 19] suggest that the combined visual channels of facial expression and body gesture are the most informative, and their integration is a mandatory process occurring early in the human processing stream. Therefore, fusing facial expression and body gesture in video sequences provides a potential way to accomplish effective affect analysis. However, there is few efforts reported on visual affect analysis by combining face and body cues [24]. Kapoor and Picard [15] presented a multi-sensor affect recognition system for classifying the affective state of interest in children who are solving puzzles, which combines the extracted sensory information from the face videos, the senor chair (body posture), and the state of the puzzle. Balomemos *et al.*[2] attempted to analyze emotions from facial expressions and hand gestures. Recently Gunes and Piccardi [11] combined expressive face and body gestures for emotion recognition in video sequences. However, how to effectively fuse these two different modalities is still an understudied problem.

In this paper, we investigate affective body gesture analysis in video sequences. Particularly, we exploit spatial-temporal features based on space-time interest point detection [6] for representing body gestures in videos. Different from the previous studies [2, 11],

which rely on much human supervision and robust hand tracking and segmentation, our approach makes few assumptions about the observed data, such as background, occlusion and appearance. The underlining motivation is that, although two instances of the body gesture representing the same emotion may vary in terms of overall appearance and motion, due to variations across subjects or within each individual, many of the spatial-temporal features detected are similar. With regard to combining different modalities, we exploit Canonical Correlation Analysis (CCA), a powerful statistical tool that is well suited for relating two sets of signals, to fuse facial expression and body gesture at the feature level. Our motivation is that, as face and body cues are two sets of measurements for affective states, conceptually the two modalities are correlated, and their relationship can be established using CCA. CCA derives a semantic "affect" space, in which the face and body features are compatible and can be effectively fused.

Compared with the previous attempts [2, 11] in vision-based affective body gesture analysis, we make the following favorable contributions: (1) we adopt spatial-temporal features based representation for body gestures, avoiding difficulties of tracking and localization in dealing with real-world video data; (2) we present to effectively fuse two modalities at the feature level using CCA; (3)we carry out study on a much large dataset.

2 Affective Body Gesture Recognition

Recently spatial-temporal features have been investigated for event detection and behavior recognition in videos [7, 18, 16, 6, 21]. Efros *et al.*[7] introduced a motion descriptor based on optical flow measurements in a spatio-temporal volume, which was applied to recognize human action at a distance. By extending 2D rectangle features into the spatio-temporal domain, Ke *et al.*[16] presented volumetric features for event detection in videos. Laptev and Lindeberg [18] extended the spatial interest points into the spatio-temporal domain, and presented a method to detect local structures in space-time where the image values have significant local variation in both space and time. Recently Dollár *et al.*[6] proposed an alternative approach to detect sparse space-time interest points based on separable linear filters, and utilized the cuboids of spatio-temporal windowed data surrounding interest points for behavior recognition. Based on their work, Niebles *et al.*[21] more recently presented an unsupervised method for action categorization. Spatial-temporal features have been proven useful to provide a compact abstract representation of video patterns. Here we adopt spatial-temporal features to represent body gesture in videos.

2.1 Spatial-Temporal Features

We extract spatial-temporal features by detecting space-time interest points in videos. Following [6, 21], we calculate the response function by application of separable linear filters. Assuming a stationary camera or a process that can account for camera motion, the response function has the form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $I(x,y,t)$ denotes images in the video, $g(x,y;\sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x,y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev}(t;\tau,\omega) =$

$-\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. In all cases we use $\omega = 4/\tau$ [6]. The two parameters σ and τ correspond roughly to the spatial and temporal scales of the detector. Each interest point is extracted as a local maxima of the response function. As pointed out in [6], any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response, while region undergoing pure translational motion, or areas without spatially distinguishing features, will not induce a strong response.

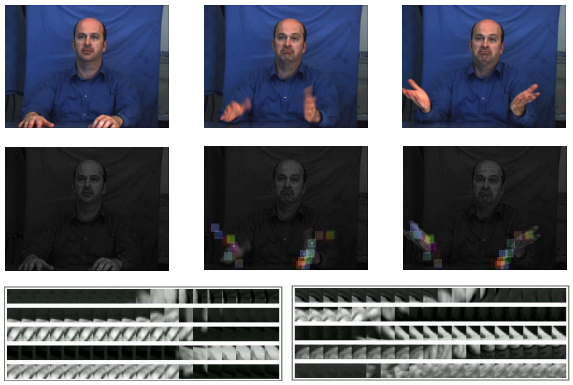

Figure 2: (Best viewed in color) Examples of spatial-temporal features extracted from videos: the first row is the original input video; the second row visualizes the cuboids extracted, where each cuboid is labeled with a different color; the third row shows some cuboids, which are flattened with respect to time.

At each detected interest point, a cuboid is extracted which contains the spatio-temporally windowed pixel values. See Figure 2 for examples of cuboids extracted. The side length of cuboids is set as approximately six times the scales along each dimension, so containing most of the volume of data that contribute to the response function at each interest point. After extracting the cuboids, the original video is discarded, which is represented as a collection of the cuboids. To compare two cuboids, different descriptors for cuboids have been evaluated in [6], including normalized pixel values, brightness gradient and windowed optical flow, followed by a conversion into a vector by flattening, global histogramming, and local histogramming. As suggested, we adopt the flattened brightness gradient as the cuboid descriptor. To reduce the dimensionality, the descriptor is projected to a lower dimensional PCA space [6]. By clustering a large number of cuboids extracted from the training data using the K-Means algorithm, we derive a library of cuboid prototypes. So each cuboid is assigned a type by mapping it to the closest prototype vector. Following [6], we use the histogram of the cuboid types to describe the video.

2.2 Recognition: SVM

we adopt the Support Vector Machine (SVM) classifier to recognize affective body gestures. SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach. SVM performs an implicit

mapping of data into a higher dimensional feature space, and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space.

Given a training set of labeled examples $\{(x_i, y_i), i = 1, \ldots, l\}$ where $x_i \in R^n$ and $y_i \in \{1, -1\}$, a new test example x is classified by the following function:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \tag{2}$$

where α_i are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(\cdot, \cdot)$ is a kernel function, and b is the threshold parameter of the hyperplane. The training sample x_i with $\alpha_i > 0$ is called the *support vector*, and SVM finds the hyperplane that maximizes the distance between the support vectors and the hyperplane. Given a non-linear mapping Φ that embeds the input data into the high dimensional space, kernels have the form of $K(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$. SVM allows domain-specific selection of the kernel function, and the most commonly used kernel functions are the linear, polynomial, and Radial Basis Function (RBF) kernels.

3 Fusing Facial Expressions and Body Gestures

A single body gesture can be ambiguous. For example, the examples shown in the second and fourth row in Figure 1 have much similar body gesture, but the affective state they express are quite different, as shown by their facial expressions. As suggested in psychological studies [1], combining visual channels of facial expression and body gesture is a potential way to accomplish effective affect analysis.

The psychological study [19] suggests that the integration of facial expression and body gesture is a mandatory process occurring early in human processing stream. So the two modalities should be processed in a joint feature space [23], rather than fused at the decision-level. The main difficulties for the feature-level fusion are the features from different modalities may be incompatible, and the relationship between different feature spaces is unknown. Here we propose to fuse face and body cues at the feature level using CCA. Our motivation is that, as facial expression and body gesture are two sets of measurements for the affective state, conceptually they are correlated. CCA can establish their relationship, deriving a semantic "affect" space, in which the face and body features are compatible and can be effectively fused.

3.1 Canonical Correlation Analysis

CCA [13] is a statistical technique developed for measuring linear relationships between two multidimensional variables. It finds pairs of base vectors (i.e., canonical factors) for two variables such that the correlations between the projections of the variables onto these canonical factors are mutually maximized. Recently CCA has been applied to computer vision problems [3, 20, 12, 17]. Borga [3] adopted CCA to find corresponding points in stereo images. Melzer *et al.*[20] applied CCA to model the relation between an object's poses with raw brightness images for appearance-based 3D pose estimation. Harsoon *et al.*[12] presented a method using CCA to learn a semantic representation to web images and their associated text.

Given two zero-mean random variables $\mathbf{x} \in R^m$ and $\mathbf{y} \in R^n$, CCA finds pairs of directions $\mathbf{w}_x$ and $\mathbf{w}_y$ that maximize the correlation between the projections $x = \mathbf{w}_x^T \mathbf{x}$ and $y = \mathbf{w}_y^T \mathbf{y}$. The projections x and y are called *canonical variates*. More formally, CCA maximizes the function:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^T \mathbf{x}\mathbf{y}^T \mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T \mathbf{x}\mathbf{x}^T \mathbf{w}_x]E[\mathbf{w}_y^T \mathbf{y}\mathbf{y}^T \mathbf{w}_y]}} = \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}} \tag{3}$$

where $\mathbf{C}_{xx} \in R^{m \times m}$ and $\mathbf{C}_{yy} \in R^{n \times n}$ are the *within-set covariance matrices* of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\mathbf{C}_{xy} \in R^{m \times n}$ denotes their *between-sets covariance matrix*. A number of at most $k = \min(m,n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ can be obtained by successively solving $\arg\max_{\mathbf{w}_x^i, \mathbf{w}_y^i} \{\rho\}$ subject to $\rho(\mathbf{w}_x^j, \mathbf{w}_x^i) = \rho(\mathbf{w}_y^j, \mathbf{w}_y^i) = 0$ for $j = 1, \ldots, i-1$, i.e., the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones.

The maximization problem can be solved by setting the derivatives of Eqn. (3), with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in the eigenvalue equations as:

$$\begin{cases} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{w}_x = \rho^2 \mathbf{w}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{w}_y = \rho^2 \mathbf{w}_y \end{cases} \tag{4}$$

Matrix inversions need to be performed in Eqn. (4), leading to numerical instability if $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are rank deficient. Alternatively, $\mathbf{w}_x$ and $\mathbf{w}_y$ can be obtained by computing principal angles, as CCA is the statistical interpretation of principal angles between two linear subspace [9] (see [17] for details).

3.2 Feature Fusion of Face and Body

Given $B = \{\mathbf{x} | \mathbf{x} \in R^m\}$ and $F = \{\mathbf{y} | \mathbf{y} \in R^n\}$, where $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors extracted from bodies and faces respectively, we apply CCA to establish the relationship between $\mathbf{x}$ and $\mathbf{y}$. Suppose $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ are the canonical factors pairs obtained, we can use d ($1 \leq d \leq k$) factor pairs to represent the correlation information. With $\mathbf{W}_x = [\mathbf{w}_x^1, \ldots, \mathbf{w}_x^d]$ and $\mathbf{W}_y = [\mathbf{w}_y^1, \ldots, \mathbf{w}_y^d]$, we project the original feature vectors as $\mathbf{x}' = \mathbf{W}_x^T \mathbf{x} = [x_1, \ldots, x_d]^T$ and $\mathbf{y}' = \mathbf{W}_y^T \mathbf{y} = [y_1, \ldots, y_d]^T$ in the lower dimensional correlation space, where x_i and y_i are uncorrelated with the previous pairs x_j and $y_j, j = 1, \ldots, i-1$. We then combine the projected feature vector $\mathbf{x}'$ and $\mathbf{y}'$ to form the new feature vector as

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x^T \mathbf{x} \\ \mathbf{W}_y^T \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x & 0 \\ 0 & \mathbf{W}_y \end{pmatrix}^T \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{5}$$

This fused feature vector effectively represents the multimodal information in a joint feature space for affect analysis.

4 Experiments

There are several facial expression databases in affective-computing community, but few databases containing affective body gestures. Gunes and Piccardi [10] recently collected a bimodal face and body gesture database (FABO), which consists of facial expression and body gesture recorded simultaneously. The database includes 23 subjects in age from

18 to 50 years, of which 12 were female, 23 were from Europe, 2 from Middle East, 3 from Latin America, 7 from Asia, and 1 from Australia. In total there are around 1900 videos. Examples of the video sequences are shown in Figure 1. In our experiments, we selected 262 videos of seven emotions (Anger, Anxiety, Boredom, Disgust, Joy, Puzzle, and Surprise) from 23 subjects. Gunes and Piccardi [11] reported some preliminary results on this database, but they only used 54 videos from 4 subjects.

4.1 Affective Body Gesture Recognition

To evaluate the algorithms' generalization ability, we adopted a 5-fold cross-validation test scheme in all recognition experiments. That is, we divided the data set randomly into five groups with roughly equal number of videos, and then used the data from four groups for training and the left group for testing; the process was repeated five times for each group in turn to be tested. We report the average recognition rates here. In all experiments, we set the soft margin C value of SVMs to infinity so that no training error was allowed. Meanwhile, each training and testing vector was scaled to be between -1 and 1. In our experiments, the RBF kernel always provided the best performance, so we report the performance of the RBF kernel. With regard to the hyper-parameter selection of RBF kernels, as suggested in [14], we carried out grid-search on the kernel parameters in the 5-fold cross-validation. The parameter setting producing the best cross-validation accuracy was picked. We used the SVM implementation in the publicly available machine learning library SPIDER [1] in our experiments.

We compare the SVM classifier with the 1-nearest neighbor classifier used in [6] for affective body gesture recognition. The average recognition rates of SVM and 1-nearest neighbor classifier are 72.6% and 68.6% respectively. We plot the confusion matrices of the two classifier in Figure 3. It can observed that the SVM classifier slightly outperforms the 1-nearest neighbor classifier.

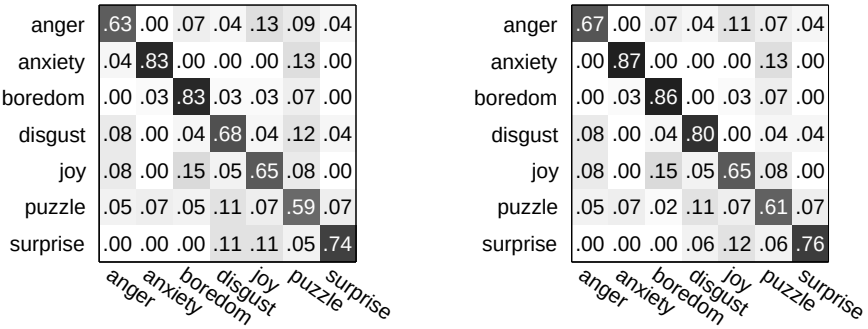

Figure 3: Confusion matrices of affective body gesture recognition with the 1-nearest neighbor classifier (*left*) and the SVM classifier (*right*).

[1] *http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html*

4.2 Emotion Recognition by Fusing Face and Body Cues

In the FABO database, video sequences were recorded simultaneously using two video cameras, one is for capturing the facial expression only and the other for capturing upper-body movements. We extracted the spatial-temporal features from the face video and the body video, and then fuse the two modalities at the feature level using CCA. We first report the classification performance based on facial cues only. The confusion matrices of the two classifiers are shown in Figure 4, and the recognition rates of SVM and 1-nearest neighbor classifier are 79.2% and 74.8% respectively. We can see that the emotion classification based on facial expressions is better than that of body gesture. This is possibly because there are much variation in affective body gestures.

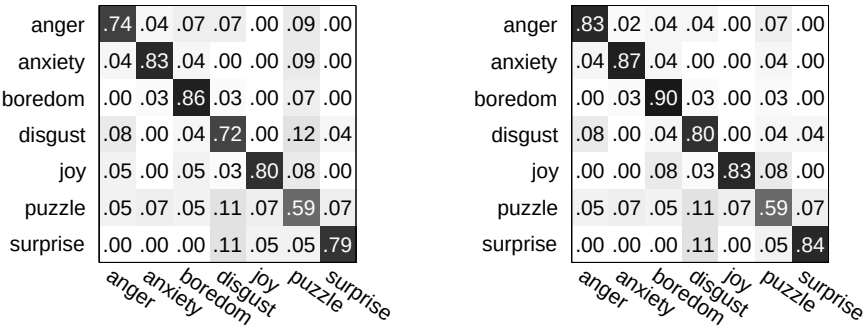

Figure 4: Confusion matrices of facial expression recognition with the 1-nearest neighbor classifier (*left*) and the SVM classifier (*right*).

We then fused facial expression and body gesture at the feature level using CCA. Different numbers of CCA factor pairs can be used to project the original face and body feature vectors to a lower dimensional CCA feature space, and the recognition performance varies with the dimensionality of the projected CCA features. We report the best result obtained here. We compared the CCA feature fusion with another three feature fusion methods: (1) Direct feature fusion, that is, concatenating the original body and face features to derive a single feature vector; (2) PCA feature fusion: the original body and face features are first projected to the PCA space respectively, and then the PCA features are concatenated to form the single feature vector. In our experiments, all principle components were kept. (3) PCA+LDA feature fusion: for each modality, the derived PCA features are further projected to the discriminant LDA space; the LDA features are then combined to derive the single feature vector. We report the experimental results of different feature fusion schemes in Table 1. The confusion matrices of the CCA feature fusion and the direct feature fusion are shown in Figure 5. We can see that the presented CCA feature fusion provides best recognition performance. This is because CCA captures the relationship between the feature sets in different modalities, and the fused CCA features effectively represent information from each modality.

Feature Fusion	CCA	Direct	PCA	PCA+LDA
Recognition Rate	88.5%	81.9%	82.3%	87.8%

Table 1: Experimental results of affect recognition by fusing body and face cues.

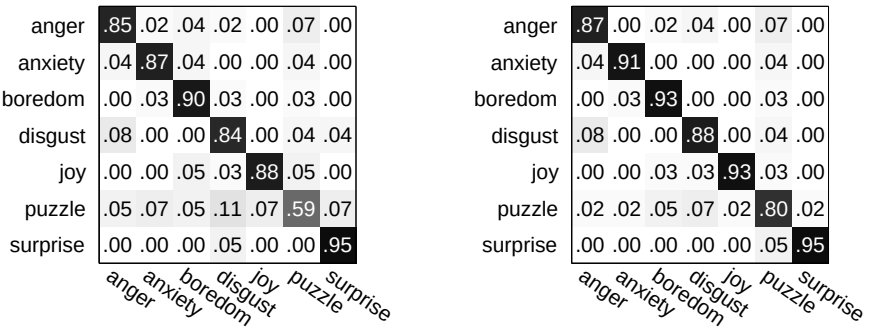

Figure 5: Confusion matrices of affect recognition by fusing facial expression and body gesture. (*left*) Direct feature fusion; (*right*) CCA feature fusion.

5 Conclusions and Discussions

In this paper, we investigate affective body gesture analysis in videos, a relatively understudied problem. Spatial-temporal features are exploited for modeling of body gestures. We also present to fuse facial expression and body gesture at the feature level using Canonical Correlation Analysis. The current spatial-temporal features based video description does not consider the position relations of cuboids detected. By including the relative position information between the cuboid types, the representation will be much more discriminative. This will be studied in our future work.

References

[1] N. Ambady and R. Rosenthal. Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, February 1992.

[2] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. In *Machine Learning for Multimodal Interaction, LNCS 3361*, pages 318–328, 2005.

[3] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linkoping University, SE-581 83 Linkoping, Sweden, 1998. Dissertation No 531.

[4] J. K. Burgoon, M. L. Jensen, T. O. Meservy, J. Kruse, and J. F. Nunamaker. Augmenting human identification of emotional states in video. In *International Conference on Intelligent Data Analysis*, 2005.

[5] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, Summer 2004.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72, 2005.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision (ICCV)*, pages 726–733, 2003.

[8] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36:259–275, 2003.

[9] G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. Technical report, 1992.

[10] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 1148–1153, 2006.

[11] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 2007.

[12] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[13] H. Hotelling. Relations between two sets of variates. *Biometrika*, 8:321–377, 1936.

[14] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Taipei, 2003.

[15] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *ACM International Conference on Multimedia*, 2005.

[16] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 166–173, 2005.

[17] T-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *European Conference on Computer Vision (ECCV)*, pages 251–262, 2006.

[18] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)*, pages 432–439, 2003.

[19] H. Meeren, C. Heijnsbergen, and B. Gelder. Rapid perceptual integration of facial expression and emotional body language. *Procedings of the National Academy of Sciences of USA*, 102(45):16518–16523, November 2005.

[20] T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 39(9):1961–1973, 2003.

[21] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference (BMVC)*.

[22] M. Pantic, A. Pentland, Nijholt A., and T.S. Huang. Human computing and machine understanding of human behavior: A survey. In T.S. Huang, A. Nijholt, M. Pantic, and A. Pentland, editors, *Artificial Intelligence for Human Computing*, volume 4451, pages 47–71. Springer, 2007.

[23] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceeding of the IEEE*, volume 91, pages 1370–1390, 2003.

[24] M. Pantic, N. Sebe, J. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *ACM International Conference on Multimedia*, pages 669–676, 2005.

[25] P. Ravindra De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterizatino of posture features. *Computer Animation and Virtual Worlds*, 15:169–276, 2004.

[26] P. Ravindra De Silva, M. Osano, and A. Marasinghe. Towards recognizing emotion with affective dimensions through body gestures. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2006.

[27] Y. Wu and T. Huang. Human hand modeling, analysis and animation in the context of human computer interaction. *IEEE Signal Processing Magazine*, 18(3):51–60, May 2001.