Visual Inference of Human Emotion and Behaviour

Shaogang Gong
Dept of Computer Science
Queen Mary College, London
England, UK
sgg@dcs.qmul.ac.uk

Caifeng Shan
Dept of Computer Science
Queen Mary College, London
England, UK
cfshan@dcs.qmul.ac.uk

Tao Xiang
Dept of Computer Science
Queen Mary College, London
England, UK
txiang@dcs.qmul.ac.uk

ABSTRACT

We address the problem of automatic interpretation of non-exaggerated human facial and body behaviours captured in video. We illustrate our approach by three examples. (1) We introduce Canonical Correlation Analysis (CCA) and Matrix Canonical Correlation Analysis (MCCA) for capturing and analyzing spatial correlations among non-adjacent facial parts for facial behaviour analysis. (2) We extend Canonical Correlation Analysis to multimodality correlation for behaviour inference using both facial and body gestures. (3) We model temporal correlation among human movement patterns in a wider space using a mixture of Multi-Observation Hidden Markov Model for human behaviour profiling and behavioural anomaly detection.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion, Perceptual reasoning, Video analysis*; I.4 [**Image Processing and Computer Vision**]: Scene Analysis—*Time-varying imagery, Object recognition*

General Terms

Algorithms, Theory

Keywords

Human emotion recognition, intention inference, body language recognition, behaviour profiling, anomaly detection

1. INTRODUCTION

To be able to visually infer automatically human behaviours is hugely desirable, not the least because its potential applications in intelligent human machine interface, healthcare and visual surveillance for public wellbeing, security and safety. There is an increasing demand for automatic methods capable of analysing human emotions and activities using video media, both for better communication and for detecting abnormal behaviours, ranging from facial behaviours, body gestures to activities in a wider space. For human behaviour analysis and inference, we advocate the need for systematic modelling of spatial and temporal correlations among facial/body parts and movement patterns that can be extracted to facilitate meaningful interpretation of behaviours. Our approach emphasises that behaviours are better interpreted in a wider spatial and temporal context. This is specially true for non-exaggerated natural and necessarily subtle behaviours.

Despite recent advance in machine perception of human emotion, much of the work remains single modality driven [28]. There has been an effort to combine images of facial expression with audio information [27]. Kapoor and Picard [18] also presented a multi-sensor affect recognition system for classifying the affective state of interest in children solving puzzles. The extracted sensory information from face videos, postures, and the state of the puzzle are combined using a Bayesian approach. Other forms of multimodal information can also be exploited for human behaviour inference. In particular, studies in psychology [1, 23] suggest that combined visual channels of facial expressions and body gestures are the most informative, and their integration is a mandatory step occurring early in human cognitive process. Both facial and body characteristics contribute holistically to conveying a more accurate emotional state of an individual. Balomemos *et al.* [2] made tentative attempt to analyze emotions from user facial expressions and hand gestures. More recently, Gunes and Piccardi [13] reported some preliminary results on combining face and body gestures for emotion recognition. In our approach, we deploy Canonical Correlation Analysis and Matrix Canonical Correlation Analysis for (1) correlating spatially non-adjacent facial parts in facial behaviour analysis, and (2) combining different modalities using both facial and body gestures.

To infer human behaviour in a wider context requires automatic behaviour profiling and discovery of underlying spatial and in particular, temporal correlation among movement patterns. In this context, we define an anomaly as an atypical behaviour pattern that is not represented by sufficient samples in a training dataset but critically it satisfies the specificity constraint to an abnormal pattern. This is because one of the main challenges for the model is to differentiate anomaly from outliers caused by noisy visual features used for behaviour representation. Much work on abnormal behaviour detection took a supervised learning approach [6, 9, 11, 24, 25] based on the assumption that there exists well-defined and known *a priori* behaviour classes

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.
Copyright 2007 ACM 978-1-59593-817-6/07/0011 ...$5.00.

(both normal and abnormal). However, natural spontaneous behaviour and any anomaly are far from being well-defined, resulting in insufficient clearly labelled data required for supervised model building. In our work, an explicit model based on a mixture of Multi-Observation Hidden Markov Model is constructed in an unsupervised manner to learn specific behaviour classes for automatic detection of abnormalities on-the-fly given unseen data. We develop a principled criterion for anomaly detection and normal behaviour recognition based on a run-time accumulative anomaly measure and an online Likelihood Ratio Test (LRT) method originally proposed for key-words detection in speech recognition [31]. This makes our approach more robust to noise in behaviour representation. This approach has two primary advantages over previous approaches, e.g. [3,14], which are: (1) it is based on constructing a generative behaviour model which scales well with the complexity of behaviour and is robust to errors in behaviour representation; (2) it performs on-the-fly anomaly detection and is therefore suitable for real-time behaviour inference.

2. FACIAL CORRELATION

Automatic facial expression and behaviour analysis has attracted much attention in recent years [26]. As facial muscles are contracted in unison to display expressions, different facial parts almost always show strong correlations. To be able to capture and analyze these correlations can facilitate better interpretation of facial behaviours. Most of the existing work on facial expression analysis [5,21,30] do not explicitly model such correlations. To this end, we employ Canonical Correlation Analysis (CCA) [16] for mapping any two sets of salient facial parts. To overcome intrinsic limitations of CCA caused by the need for matrix to vector concatenation, we further develop a novel Matrix Canonical Correlation Analysis (MCCA) for correlation analysis of images in their native 2D array form. Our experiments demonstrate that MCCA can better measure correlations in 2D image data, providing superior performance in regression and recognition tasks, whilst requiring much fewer canonical factors.

2.1 Canonical Correlation Analysis

CCA was developed originally by Hotelling [16] for measuring linear relationships between two vector variables. It finds pairs of base vectors, i.e. canonical factors, for two variables such that the correlations between the projections of these variables onto the canonical factors are mutually maximised. Similar to Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), CCA also reduces the dimensionality of the original data. Whilst PCA aims to minimise the reconstruction error and LDA aims to both maximise between-class scatter and minimise within-class scatter, CCA seeks for directions of two sets of vectors that maximise their inter-correlation. Recently CCA has been exploited for solving computer vision and pattern recognition problems [4,8,15,19].

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 maximise 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 maximises the following function:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^T\mathbf{x}\mathbf{y}^T\mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T\mathbf{x}\mathbf{x}^T\mathbf{w}_x]E[\mathbf{w}_y^T\mathbf{y}\mathbf{y}^T\mathbf{w}_y]}}$$

$$= \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}} \tag{1}$$

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 maximum of $k = \min(m,n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1,\dots,k$, 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 $j = 1,\dots,i-1$, i.e. the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones. Canonical variates x_i and y_i (corresponding to $\mathbf{w}_x^i$ and $\mathbf{w}_y^i$) are uncorrelated with the previous pairs x_j and $y_j, j = 1,\dots,i-1$.

This maximisation problem can be solved by setting the derivatives of Eqn. (1), with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in the following eigenvalue equations:

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

Matrix inversions need to be performed in Eqn. (2), 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 [10].

2.2 Matrix Canonical Correlation Analysis

Existing CCA shares a number of problems with other subspace analysis methods such as PCA and LDA. Applying CCA to images requires two-dimensional image arrays concatenated into one-dimensional vectors. This matrix-to-vector operation leads to two main problems. Firstly, the intrinsic 2D structure of image matrices is removed, so the spatial information stored therein is discarded. CCA based on these vectors cannot fully capture correlations among the original 2D image data. Secondly, each image sample is modeled as a high-dimensional vector so that a large number of training samples are needed to yield a reliable estimation of the underlying data distribution. In reality, very limited number of training data are usually available. To address these problems, we introduce a novel Matrix Canonical Correlation Analysis (MCCA) for correlation analysis of images in their native 2D array form.

Given two matrix variables $\mathbf{A} \in R^{m\times n}$ and $\mathbf{B} \in R^{j\times k}$ (we assume the variables are both zero-mean), MCCA finds pairs of directions $\mathbf{v}_a \in R^m$, $\mathbf{w}_a \in R^n$, $\mathbf{v}_b \in R^j$ and $\mathbf{w}_b \in R^k$ that maximise the correlation between the projections $a = \mathbf{v}_a^T\mathbf{A}\mathbf{w}_a$ and $b = \mathbf{v}_b^T\mathbf{B}\mathbf{w}_b$. Mathematically, we can formulate this as the following maximisation problem: Find optimal $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ that maximise

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^T\mathbf{A}\mathbf{w}_a\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T\mathbf{A}\mathbf{w}_a\mathbf{w}_a^T\mathbf{A}^T\mathbf{v}_a]E[\mathbf{v}_b^T\mathbf{B}\mathbf{w}_b\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b]}} \tag{3}$$

Here $\mathbf{v}_a$ ($\mathbf{v}_b$) and $\mathbf{w}_a$ ($\mathbf{w}_b$) are canonical factors in two dimensions, acting as a two-sided linear transformation on the data in matrix form. There is no closed-form solution for the maximisation problem in Eqn. (3). Alternatively, we present an iterative procedure for computing $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ as follows: Given an initial choice of $\mathbf{w}_a$ and $\mathbf{w}_b$, we estimate $\mathbf{v}_a$ and $\mathbf{v}_b$ by computing canonical factors of $\mathbf{a}'$ and $\mathbf{b}'$; with $\mathbf{v}_a$ and $\mathbf{v}_b$ (corresponding to the largest canonical correlation), we then estimate $\mathbf{w}_a$ and $\mathbf{w}_b$ by computing canonical factors of $\mathbf{a}''$ and $\mathbf{b}''$; $\mathbf{w}_a$ and $\mathbf{w}_b$ (corresponding to the largest canonical correlation) are used iteratively. This procedure is repeated until convergence such that a maximum of $q = \min(m, j)$ left-side canonical factor pairs $\langle \mathbf{v}_a^1, \mathbf{v}_b^1 \rangle, \ldots, \langle \mathbf{v}_a^q, \mathbf{v}_b^q \rangle$ and $p = \min(n, k)$ right-side canonical factor pairs $\langle \mathbf{w}_a^1, \mathbf{w}_b^1 \rangle, \ldots, \langle \mathbf{w}_a^p, \mathbf{w}_b^p \rangle$ are computed.

3. MULTIMODAL CORRELATION

Single mode facial/body gestures are often highly ambiguous. Psychological studies [23] suggest that interpreting facial expression and body gesture together is a mandatory step occurring early in human cognitive process. In this context, we are interested in modelling/discovering any correlation between facial and body gestures at the feature representational level using CCA. For feature representation, we employ spatial-temporal features based on space-time interest point detection in video [7]. Previous studies [2, 13], where tracked hand motion was applied to gesture recognition, rely upon intensive human supervision and are based on assumptions that reliable hand tracking and segmentation can be achieved therefore stable background with minimal occlusion and appearance change are required. Our approach makes no such assumptions about the observed video data. Our experiments show that despite two instances of the same body gestures may change in both appearance and motion, due to variations across subjects, or within each individual, the detected space-time interest point features are stable.

More precisely, given $F = \{\mathbf{x} | \mathbf{x} \in R^m\}$ and $B = \{\mathbf{y} | \mathbf{y} \in R^n\}$, where $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors extracted from face and body 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. We then combine the projected feature vector $\mathbf{x}'$ and $\mathbf{y}'$ to form a new feature vector

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x^T \mathbf{x} \\ \mathbf{W}_y^T \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x & 0 \\ 0 & \mathbf{W}_y \end{pmatrix}^T \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \qquad (4)$$

This fused feature vector effectively represents the bi-modal information for emotion recognition.

4. BEHAVIOUR PROFILING

We define behaviour profiling as follows. Consider a training dataset $\mathbf{D}$ consisting of N feature vectors:
$\mathbf{D} = \{\mathbf{P}_1, \ldots, \mathbf{P}_n, \ldots, \mathbf{P}_N\}$, where $\mathbf{P}_n$ is a set of features [32] extracted from successive image frames representing a behaviour pattern captured by the nth video segment $\mathbf{v}_n$. The problem to be addressed is to discover the natural grouping/correlation of given training behaviour patterns upon which a model for normal behaviour can be built. This is essentially a data clustering problem with the number of clusters unknown. There are a number of aspects that make this problem challenging: (1) Each feature vector $\mathbf{P}_n$ can be of different lengths. Conventional clustering approaches such as K-means and mixture models require that each data sample is represented as a fixed length feature vector. These approaches thus cannot be applied directly. (2) A definition of a distance/affinity metric among these variable length feature vectors is nontrivial. Measuring affinity between feature vectors of variable length often involves Dynamic Time Warping [20]. A standard dynamic time warping (DTW) method used in computer vision community would attempt to treat the feature vector $\mathbf{P}_n$ as a K_e dimensional trajectory and measure the distance of two behaviour patterns by finding correspondence between discrete vertices on two trajectories. Since in our framework, a behaviour pattern is represented as a set of temporal correlated visual features, i.e. a stochastic process, a stochastic modelling based approach is more appropriate for distance measuring. Note that in the case of matching two sequences of different lengths based on video object detection, the affinity of the most similar pair of images from two sequences can be used for sequence affinity measurement [29]. However, since we focus on modelling behaviour that could involve multiple objects interacting over space and time, the approach in [29] can not be applied directly in our case. (3) Model selection needs performed to determine the number of clusters. To overcome these difficulties, we propose a spectral clustering algorithm with feature and model selection based on modelling each behaviour pattern using a Dynamic Bayesian Network.

4.1 Affinity Matrix

Dynamic Bayesian Networks (DBNs) provide a solution for measuring the affinity between different behaviour patterns. More specifically, each behaviour pattern in the training set is modelled using a DBN. To measure the affinity between two behaviour patterns represented as $\mathbf{P}_i$ and $\mathbf{P}_j$, two DBNs denoted as $\mathbf{B}_i$ and $\mathbf{B}_j$ are trained on $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively using the EM algorithm. The affinity between $\mathbf{P}_i$ and $\mathbf{P}_j$ is then computed as:

$$S_{ij} = \frac{1}{2} \left\{ \frac{1}{T_j} \log P(\mathbf{P}_j | \mathbf{B}_i) + \frac{1}{T_i} \log P(\mathbf{P}_i | \mathbf{B}_j) \right\}, \qquad (5)$$

where $P(\mathbf{P}_j | \mathbf{B}_i)$ is the likelihood of observing $\mathbf{P}_j$ given $\mathbf{B}_i$, and T_i and T_j are the lengths of $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively.

DBNs of different topologies can be employed. A straightforward choice would be a Hidden Markov Model (HMM). However, a drawback of a HMM is that too many parameters are needed to describe the model when the observation variables are of high dimension. This makes a HMM vulnerable to overfitting therefore generating poorly to unseen data. It is especially true in our case because a HMM needs to be learned for every single behaviour pattern in the training dataset which could be short in duration. To solve this problem, we employ a Multi-Observation Hidden Markov Model (MOHMM) [11]. Compared to a HMM, the observational space is factorised by assuming that each observed feature (p_{nt}^k) is independent of each other. Consequently, the number of parameters for describing a MOHMM is much lower than that for a HMM.

An $N \times N$ affinity matrix $\mathbf{S} = [S_{ij}]$ where $1 \leq i, j \leq N$ provides a new representation for the training dataset, denoted as $\mathbf{D_s}$. In this representation, a behaviour pattern is represented by its affinity to each behaviour pattern in the training set. Specifically, the nth behaviour pattern is now represented as the nth row of $\mathbf{S}$, denoted as $\mathbf{s}_n$. The natural grouping of behaviour patterns in the training data set is then discovered through a novel spectral clustering algorithm which measure the relevance of each eigenvector and performs clustering using only the selected relevant eigenvectors of the data affinity matrix [32].

4.2 Behaviour as A Mixture of MOHMMs

To build a generative model for the observed/expected behaviour, we first model the kth behaviour class using a MOHMM $\mathbf{B}_k$. The parameters of $\mathbf{B}_k$, $\theta_{\mathbf{B}_k}$ are estimated using all the patterns in the training set that belong to the kth class. A behaviour model $\mathbf{M}$ is then formulated as a mixture of the K MOHMMs. Given an unseen behaviour pattern, represented as a behaviour pattern feature vector $\mathbf{P}$, the likelihood of observing $\mathbf{P}$ given $\mathbf{M}$ is

$$P(\mathbf{P}|\mathbf{M}) = \sum_{k=1}^{K} \frac{N_k}{N} P(\mathbf{P}|\mathbf{B}_k), \tag{6}$$

where N is the total number of training behaviour patterns and N_k is the number of patterns that belong to the kth behaviour class.

4.3 Online Anomaly Detection

Once constructed, the generative behaviour model $\mathbf{M}$ can be used to detect whether an unseen behaviour pattern is normal using a run-time anomaly measure.

An unseen behaviour pattern of length T is represented as $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_t, \ldots, \mathbf{p}_T]$. At the tth frame, the accumulated visual information for the behaviour pattern, represented as $\mathbf{P}_t = [\mathbf{p}_1, \ldots, \mathbf{p}_t]$, is used for online reliable anomaly detection. First the normalised log-likelihood of observing $\mathbf{P}$ at the tth frame given the behaviour model $\mathbf{M}$ is computed as

$$l_t = \frac{1}{t} \log P(\mathbf{P}_t|\mathbf{M}). \tag{7}$$

l_t can be easily computed online using the forward-backward procedure [22]. Note that the complexity of computing l_t is $\mathcal{O}({K_e}^2)$ and does not increase with t. We then measure the anomaly of $\mathbf{P}_t$ using an online anomaly measure Q_t:

$$Q_t = \begin{cases} l_1 & \text{if } t = 1 \\ (1-\alpha)Q_{t-1} + \alpha(l_t - l_{t-1}) & \text{otherwise} \end{cases} \tag{8}$$

where α is an accumulating factor determining how important the visual information extracted from the current frame is for anomaly detection. We have $0 < \alpha \leq 1$. Compared to l_t as an indicator of normality/anomaly, Q_t could add more weight to more recent observations. Anomaly is detected at frame t if

$$Q_t < Th_A \tag{9}$$

where Th_A is the anomaly detection threshold. Note that it takes a time delay for Q_t to stabilise at the beginning of evaluating a behaviour pattern due to the nature of the forward-backward procedure. The length of this time period, denoted as T_w is related to the complexity of the MOHMM. We thus set $T_w = 3K_e$ in our experiments to be reported later in Section 5, i.e. the anomaly of a behaviour pattern is only evaluated when $t > T_w$.

5. EXPERIMENTS

5.1 Facial Correlation Analysis

As a case study, we investigate correlations between the mouth part (Mouth) and the right eye part (Eye) (as shown in Figure 1). These two parts have strong and a range of correlations corresponding to facial expressions. We conducted experiments on the Cohn-Kanade database [17] and face expression image sequences we captured. We manually normalized the faces based on three feature points, centers of the two eyes and the mouth, using affine transformation. In the normalized facial images (110×150 pixels), the mouth part is 53×68 pixels, and the eye part is 45×51 pixels.

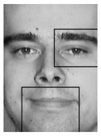


Figure 1: A case study on correlations between the mouth and the right eye facial parts.

5.1.1 Facial Parts Synthesis

We wish to reconstruct (synthesize) Mouth from Eye or vice versa using MCCA based regression. Specifically, to reconstruct image $\mathbf{B}$ from image $\mathbf{A}$, we first employ MCCA to establish their relationship, finding optimal projection directions in the sense of correlation, and then map A to the leading canonical variates by discarding directions with low canonical correlation. Finally we perform regression of $\mathbf{B}$ by taking these leading canonical variates of $\mathbf{A}$. The procedure of synthesis is as follows.

1. Compute the leading factor pairs $\mathbf{V}_a, \mathbf{W}_a, \mathbf{V}_b, \mathbf{W}_b$ from N pairs of samples $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N\}$ and $\mathcal{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_N\}$.
2. Map $\mathbf{A}_i$ $(i = 1, \ldots, N)$ to the reduced correlation space $\widetilde{\mathbf{A}}_i = \mathbf{V}_a^T \mathbf{A}_i \mathbf{W}_a$.
3. Reshape 2D matrices $\widetilde{\mathbf{A}}_i$ and $\mathbf{B}_i$ to 1D vectors $\tilde{\mathbf{a}}_i$ and $\mathbf{b}_i$, and form data matrices $\widetilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_N]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$; then compute the regression matrix $\mathbf{R} = (\widetilde{\mathbf{A}}^T)^{-1}\mathbf{B}^T$.
4. Given a new input $\mathbf{A}_{new}$, the corresponding $\mathbf{B}_{new}$ is reconstructed by:

$$\widetilde{\mathbf{A}}_{new} = \mathbf{V}_a^T \mathbf{A}_{new} \mathbf{W}_a, \quad \widetilde{\mathbf{A}}_{new} \to \tilde{\mathbf{a}}_{new} \tag{10}$$

$$\mathbf{b}_{new} = \mathbf{R}^T \tilde{\mathbf{a}}_{new}, \quad \mathbf{b}_{new} \to \mathbf{B}_{new} \tag{11}$$

Here $\widetilde{\mathbf{A}}_{new} \to \tilde{\mathbf{a}}_{new}$ represents reshaping 2D matrix $\widetilde{\mathbf{A}}_{new}$ to 1D vector $\tilde{\mathbf{a}}_{new}$, and $\mathbf{b}_{new} \to \mathbf{B}_{new}$ is reshaping 1D vector $\mathbf{b}_{new}$ to 2D matrix $\mathbf{B}_{new}$.

We selected more than 10 subjects from the Cohn-Kanade database, each of which has around 70∼140 images of different facial expressions, in addition to the image sequences we captured. For the image set of each subject, we randomly sampled one tenth of the images as the testing set, and the remaining images as the training set. We applied MCCA, CCA, and the standard linear least-squares regression (SR) approach to synthesize Mouth from Eye and vice versa on the testing set. We used 10 randomly selected training/testing combinations for reporting reconstruction errors. We observe that MCCA performs better than CCA and SR in reconstructing one facial part from another. Moreover, MCCA requires much fewer canonical factors to obtain better reconstruction results. Reconstruction results for six randomly selected subjects are shown in Table 1, with optimal average pixel errors (with standard deviation) and corresponding dimensions of canonical factors used. Some reconstruction examples are shown in Figure 2. It is evident

		Subject (1) Errors (Dims)	Subject (2) Errors (Dims)	Subject (3) Errors (Dims)	Subject (4) Errors (Dims)	Subject (5) Errors (Dims)	Subject (6) Errors (Dims)
Eye	MCCA	11.2±2.0 (66)	8.5±2.5 (54)	13.4±5.5 (45)	16.3±5.0 (39)	9.9±1.8 (28)	13.8±2.4 (28)
↓	CCA	16.7±4.4 (139)	13.0±6.0 (119)	16.2±8.8 (96)	24.5±9.8 (96)	12.9±4.4 (85)	17.2±8.5 (77)
Mouth	SR	17.3±3.5 (-)	12.4±5.3 (-)	16.1±8.8 (-)	23.7±7.6 (-)	14.2±4.3 (-)	15.1±4.9 (-)
Mouth	MCCA	8.8±1.2 (46)	8.4±1.8 (50)	10.0±3.3 (28)	19.5±6.0 (51)	10.5±2.5 (44)	12.6±2.9 (36)
↓	CCA	13.1±4.0 (139)	10.7±3.6 (119)	11.6±5.9 (96)	25.4±18.3 (96)	11.0±3.1 (85)	15.7±6.2 (77)
Eye	SR	14.7±4.8 (-)	10.7±3.2 (-)	12.0±6.5 (-)	26.1±18.8 (-)	11.0±2.9 (-)	13.9±6.1 (-)

Table 1: Reconstruction results for six subjects: the optimal average pixel errors (with standard deviation) of the three algorithms, and the corresponding dimensions of canonical factors used in MCCA and CCA.

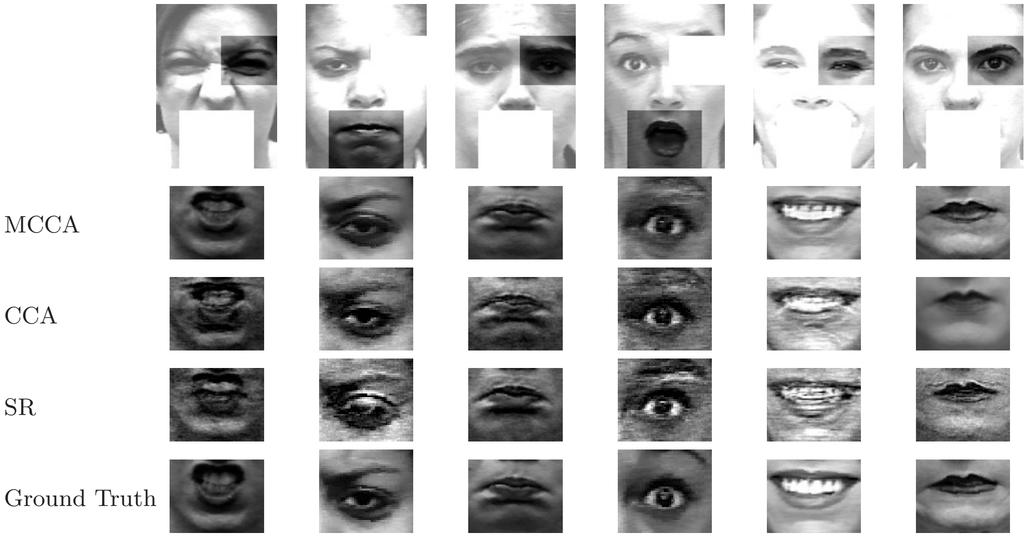


Figure 2: Some examples of facial parts synthesis using MCCA, CCA, and SR.

that MCCA outperforms CCA and SR consistently in facial parts synthesis. Crucially, the dimension of canonical factors needed in MCCA is always less than 50% of that of CCA. So MCCA can describe correlations among facial parts with better accuracy using much less canonical factors. The strength of MCCA is also reflected by the average standard deviation. As shown in Table 1, MCCA always produces the smallest deviation, which suggests that MCCA is much more robust.

5.1.2 *Expression Recognition by Correlation*

We also carried out experiments on facial expression recognition solely based on correlations between Mouth and Eye. Given image sets of different facial expressions $I_1, \ldots, I_c$ (c is the number of classes), we derive the leading factor pairs $(\mathbf{V}_a^i, \mathbf{W}_a^i, \mathbf{V}_b^i, \mathbf{W}_b^i), i = 1 \ldots c$ of parts Mouth (denoted by $\mathbf{B}$) and Eye (denoted by $\mathbf{A}$) for each class using MCCA. We then compute the regression parameters for reconstructing $\mathbf{B}$ from $\mathbf{A}$ in the reduced correlation space in the training set. Given a test image $\mathbf{I}_{new}$ of an unknown class, we map its Eye $\mathbf{A}_{new}$ and Mouth $\mathbf{B}_{new}$ to the reduced correlation space of class i as $\widetilde{\mathbf{A}}_i = (\mathbf{V}_a^i)^T \mathbf{A}_{new} \mathbf{W}_a^i$ and $\widetilde{\mathbf{B}}_i = (\mathbf{V}_b^i)^T \mathbf{B}_{new} \mathbf{W}_b^i$, and then calculate the error $err(i)$ of reconstructing $\widetilde{\mathbf{B}}_i$ from $\widetilde{\mathbf{A}}_i$ with the regression parameters of this correlation space. After computing the reconstruction error of each class $err(i), i = 1 \ldots c$, we classify the test image as the class having the smallest reconstruction error

$$\hat{i} = \arg\min_i err(i) \tag{12}$$

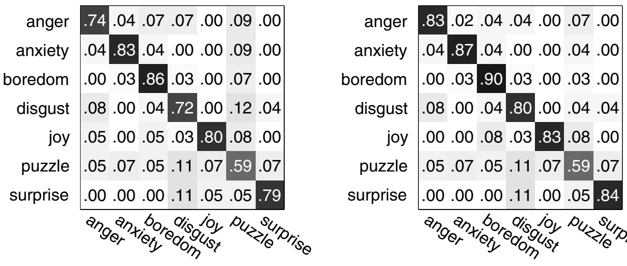
	2-Class	3-Class	4-Class
MCCA	96.1±3.6	80.8±6.4	67.9±4.8
CCA	63.2±10.5	55.6±7.8	48.7±6.7

Table 2: Facial expression recognition based on correlations of Mouth and Eye modeled by MCCA and CCA.

For our experiments, we selected 732 image of basic emotions (Anger, Disgust, Joy, and Surprise) from the Cohn-Kanade database. The sequences come from 96 subjects, with 1 to 4 emotions per subject. We first considered a 2-class (Joy and Surprise) recognition problem, then included Anger for a 3-class problem, and finally considered four expressions for classification (incrementally making the recognition task harder). To evaluate generalization performance, a 10-fold Cross-Validation testing scheme was adopted. The recognition results using MCCA and CCA are shown in Table 2. It is evident that expressions can be better classified using MCCA, demonstrating again that MCCA outperform CCA in capturing correlations in facial parts. It is also evident that by modeling correlations between only two facial parts, the recognition accuracy degrades quickly for multi-class recognition. By considering correlations of multiple facial parts, we should be able to improve these recognition results.

5.2 Multimodal Emotion Recognition

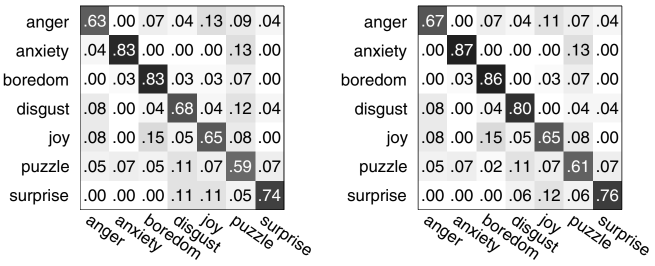
Gunes and Piccardi [12] collected the first bimodal face and body gesture database (FABO), in which video sequences were recorded simultaneously using two cameras, one capturing the head and the other capturing upper-body movements. The database includes 23 subjects aged from 18 to 50. In total there are 1900 videos. In our experiments, we selected 262 videos of seven emotions (Anger, Anxiety, Boredom, Disgust, Joy, Puzzle, and Surprise) from 23 subjects. Gunes and Piccardi [13] reported results using a smaller set of 54 videos from 4 subjects.



	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.74	.04	.07	.07	.00	.09	.00
anxiety	.04	.83	.04	.00	.00	.09	.00
boredom	.00	.03	.86	.03	.00	.07	.00
disgust	.08	.00	.04	.72	.00	.12	.04
joy	.05	.00	.05	.03	.80	.08	.00
puzzle	.05	.07	.05	.11	.07	.59	.07
surprise	.00	.00	.00	.11	.05	.05	.79

	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.83	.02	.04	.04	.00	.07	.00
anxiety	.04	.87	.04	.00	.00	.04	.00
boredom	.00	.03	.90	.03	.00	.03	.00
disgust	.08	.00	.04	.80	.00	.04	.04
joy	.00	.00	.08	.03	.83	.08	.00
puzzle	.05	.07	.05	.11	.07	.59	.07
surprise	.00	.00	.00	.11	.00	.05	.84

Figure 3: Confusion matrices of facial expression recognition using (*left*) 1-nearest neighbor classifier and (*right*) SVM classifier.



	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.63	.00	.07	.04	.13	.09	.04
anxiety	.04	.83	.00	.00	.00	.13	.00
boredom	.00	.03	.83	.03	.03	.07	.00
disgust	.08	.00	.04	.68	.04	.12	.04
joy	.08	.00	.15	.05	.65	.08	.00
puzzle	.05	.07	.05	.11	.07	.59	.07
surprise	.00	.00	.00	.11	.11	.05	.74

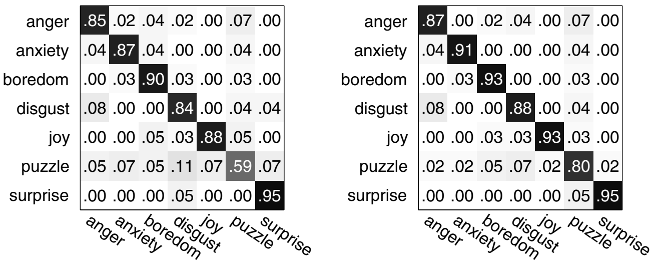
	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.67	.00	.07	.04	.11	.07	.04
anxiety	.00	.87	.00	.00	.00	.13	.00
boredom	.00	.03	.86	.00	.03	.07	.00
disgust	.08	.00	.04	.80	.00	.04	.04
joy	.08	.00	.15	.05	.65	.08	.00
puzzle	.05	.07	.02	.11	.07	.61	.07
surprise	.00	.00	.00	.06	.12	.06	.76

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

In our experiment, we first show classification performance using single modality only (facial or body features alone). The confusion matrices from two classifiers are shown in Figures 3 and 4. The error rates of SVM and 1-nearest neighbor classification of facial expression are 20.8% and 25.2% respectively, and the corresponding error rates in classifying body expression are 27.4% and 31.4% respectively. For multimodality recognition, we then extracted the spatial-temporal features from the face video and the body video simultaneously, and then fuse the two modalities at the feature level using CCA. For comparison, we also carried out experiments using direct feature fusion, i.e. concatenating the original face and body feature vectors to derive a single feature vector. Our results are shown in Figure 5, where the average error rates of CCA-based feature fusion and the direct feature fusion are 11.5% and 18.1% respectively. This shows whilst a direct feature fusion by concatenation yields only slight performance improvement over a single modality, a multimodality correlated representation using CCA feature fusion provides much improved recognition rate. This is because CCA captures underlying relationship between the feature sets in different modality spaces.

5.3 Behaviour Profiling / Anomaly Detection

A CCTV camera was mounted on the ceiling of an office entrance/exit corridor, monitoring people entering and leaving an office area (see Figure 6). The office area is secured by an entrance-door which can only be opened by scanning an entry card on the wall next to the door (see middle frame in row (b) of Figure 6). Two side-doors were also located at the right hand side of the corridor. People from both inside and outside the office area have access to those two side-doors. Typical behaviour occurring in the scene would be people entering or leaving either the office area or the side-doors, and walking towards the camera. Each behaviour pattern would normally last a few seconds. For this experiment, a dataset was collected over 5 different days consisting of 6 hours of video totalling 432000 frames captured at 20Hz with 320×240 pixels per frame. This dataset was then segmented into sections separated by any motionless intervals lasting for more than 30 frames. This resulted in 142 video segments of actual behaviour pattern instances. Each segment has on average 121 frames with the shortest 42 and longest 394.



	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.85	.02	.04	.02	.00	.07	.00
anxiety	.04	.87	.04	.00	.00	.04	.00
boredom	.00	.03	.90	.03	.00	.03	.00
disgust	.08	.00	.00	.84	.00	.04	.04
joy	.00	.00	.05	.03	.88	.05	.00
puzzle	.05	.07	.05	.11	.07	.59	.07
surprise	.00	.00	.00	.05	.00	.00	.95

	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.87	.00	.02	.04	.00	.07	.00
anxiety	.04	.91	.00	.00	.00	.04	.00
boredom	.00	.03	.93	.00	.00	.03	.00
disgust	.08	.00	.00	.88	.00	.04	.00
joy	.00	.00	.03	.03	.93	.03	.00
puzzle	.02	.02	.05	.07	.02	.80	.02
surprise	.00	.00	.00	.00	.00	.05	.95

Figure 5: Confusion matrices of fusing facial expressions and body gestures recognition. (*left*) Direct feature fusion; (*right*) CCA-based feature fusion.

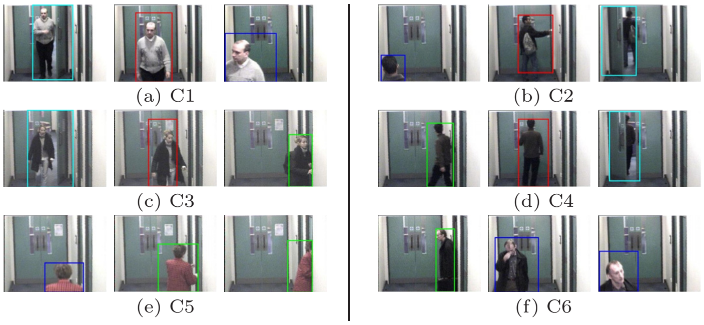


Figure 6: Behaviour patterns in a corridor entrance/exit scene. (a)–(f) show image frames of typical behaviour patterns belonging to the 6 behaviour classes listed in Table 3. The four classes of visual events detected automatically, 'entering/leaving the near end of the corridor', 'entering/leaving the entry-door', 'entering/leaving the side-doors', and 'in corridor with the entry door closed', are highlighted in the image frames using bounding boxes in blue, cyan, green and red respectively. The same colour scheme will be used for illustrating detected events in Figure 7.

C1	From the office area to the near end of the corridor
C2	From the near end of the corridor to the office area
C3	From the office area to the side-doors
C4	From the side-doors to the office area
C5	From the near end of the corridor to the side-doors
C6	From the side-doors to the near end of the corridor

Table 3: The 6 classes of behaviour patterns that most commonly occurred in a corridor entrance/exit scene.

A training set consisting of 80 video segments was randomly selected from the overall 142 segments without any behaviour class labelling of the video segments. The remaining 62 segments were used for testing the trained model later. This model training exercise was repeated 20 times and in each trial a different model was trained using a different random training set. This is in order to avoid any bias in the anomaly detection and normal behaviour recognition results to be discussed later. For these unlabelled model training over the 20 trials, the number of clusters for each training set was determined automatically as 6 in every trial. By observation, each discovered data cluster mainly contained samples corresponding to one of the 6 behaviour classes listed in Table 3. For each unlabelled training set, a normal behaviour model was constructed as a mixture of 6 MOHMMs as described in Section 4.2.

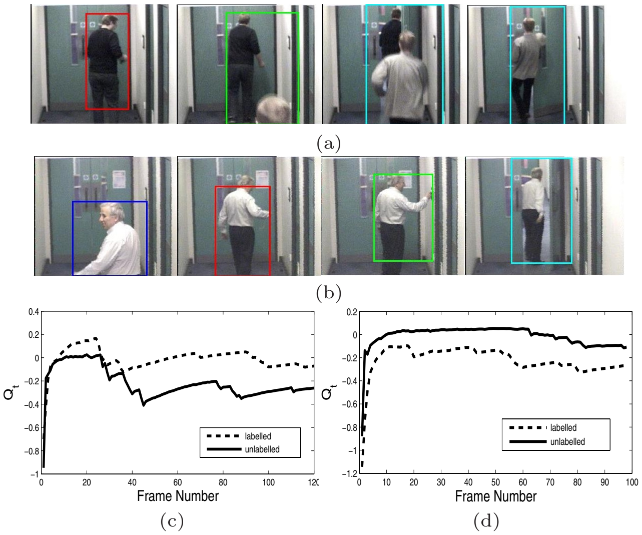
For comparative evaluation, alternative models were also trained using labelled datasets as follows. For each of the 20 training sessions above, a model was trained using identical training sets as above. However, each behaviour pattern in the training sets was also manually labelled as one of the manually identified behaviour classes. On average 12 behaviour classes were manually identified for the labelled training sets in each trial. Six classes were always identified in each training set (see Table 3). On average they accounted for 83% of the labelled training data. A normal behaviour model was built as a mixture of MOHMMs with the number of mixture components determined by the number of manually identified behaviour classes. Each MOHMM component was trained using the data samples corresponding to one class of manually identified behaviour in each training set.

Given a training set, discrete visual events were detected and classified using automatic model order selection in clustering, resulting in four classes of events corresponding to the common constituents of all behaviour in this scene: 'entering/leaving the near end of the corridor', 'entering/leaving the entrance-door', 'entering/leaving the side-doors', and 'in corridor with the entrance-door closed'. Examples of detected events are shown in Figure 6 using colour-coded bounding boxes. It is evident that that due to the narrow view nature of the scene, differences between the four common events are rather subtle and can be mis-identified based on local information (space and time) alone, resulting in large error margin in event detection. These events being also common constituents to different behaviour patterns reinforces the assumption that local events treated in isolation hold little discriminative information for behaviour profiling.

	Anomaly Det. (%)	Fal. Alarm (%)
Unlabelled	85.4 ± 2.9	6.1 ± 3.1
Labelled	73.1 ± 12.9	8.4 ± 5.3

Table 4: The mean and standard deviation of the anomaly detection rate and false alarm rates for corridor entrance/exit behaviour models trained using unlabelled and labelled data. The results were obtained over 20 trials with $Th_A = -0.2$.

The behaviour models built using both labelled and unlabelled behaviour patterns were used to perform online anomaly detection. To measure the performance of the learned models on anomaly detection, each behaviour pattern in the testing sets was manually labelled as normal if there were similar patterns in the corresponding training sets and abnormal otherwise. A testing pattern was detected as being abnormal when Eqn. (9) was satisfied at any time after $T_w = 3K_e = 12$ frames. The accumulating factor α for computing Q_t was set to 0.1. We measure the performance of anomaly detection using anomaly detection rate and false alarm rate. The detection rate and false alarm rate of anomaly detection are shown in Table 4. This suggests that the models trained using unlabelled data clearly outperformed those trained using labelled data. In particular, it is found that given the same Th_A the models trained using unlabelled datasets achieved higher anomaly detection rate and lower false alarm rate compared to those trained using labelled datasets. Figure 7 shows examples of false alarm and mis-detection by models trained using labelled data. The lower tolerance towards event detection error was the main reason for the higher false alarm rate of models trained using labelled data (see Figure 7(b)&(d) for an example).



(a)

(b)

(c)

(d)

Figure 7: Examples of anomaly detection in the corridor entrance/exit scene. (a)&(c): An abnormal behaviour pattern was detected as being abnormal by the model trained using an unlabelled dataset, while it was detected as being normal by the model trained using the same but labelled dataset. It shows a person sneaking into the office area without using an entry card. (b)&(d): A normal behaviour pattern which was detected correctly by the model trained using an unlabelled dataset, but detected as being abnormal by the model trained using the same but labelled dataset. The third frame from left in (b) shows an error in event detection (an 'in corridor with the entrance-door closed' event was detected as an 'entering/leaving the side-doors' event). Note that a smaller value of Q_t means that it is more likely for the behaviour pattern to be abnormal. Th_A was set to -0.2.

6. CONCLUSION

In summary, we presented an approach to systematic modelling of spatial and temporal correlations among facial/body parts and movement patterns for facilitating meaningful interpretation of human behaviour. Our approach emphasises that behaviour is better interpreted in a wider spatial and temporal context. This is specially true for non-exaggerated natural behaviours. In particular, we introduced Canonical Correlation Analysis and Matrix Canonical Correlation Analysis for capturing and analyzing spatial correlations among non-adjacent facial parts for facial behaviour analysis. As facial muscles are contracted in unison to display expressions, different facial parts almost always show strong correlations. To be able to capture and analyze these correlations can facilitate better interpretation of facial behaviour. Moreover, for visual interpretation of human emotion, both facial and body characteristics contribute holistically to conveying a more accurate emotional state of a person. To this end, we investigated ways to correlate multimodal visual features for more holistic and reliable emotion recognition. To improve model sensitivity, we are currently investigating the effects of correlating multiple facial and body components in space and over time. In a wider context, we further developed an approach for robust human behaviour profiling and anomaly detection using temporal correlation of auto-discovered multiple visual features and associated events. The framework is fully unsupervised. The effectiveness and robustness of our approach is demonstrated through experiments using datasets collected from natural public scenes where spontaneous behaviours are monitored.

To conclude, despite the best efforts from an increasing number of researchers, we are still at the very early stage of quantifying human behaviour, especially spontaneous and natural behaviours. Many basic questions remain to be answered. One of the questions is how to extract and make use of cross-modality domain knowledge. How can a machine discover inherent common structures and trends hidden in multimodal data exhibiting different apparent characteristics? How information in one domain can be mapped to others? Human cognitive process has the ability to associate and generalise observations across domain and modality, so that visually subtle behaviour changes can be detected and interpreted in the appropriate context. How to enable a machine to possess the same ability? We envisage that these questions will captivate researchers for years to come.

7. 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] O. Boiman and M. Irani. Detecting irregularities in images and video. In *IEEE International Conference on Computer Vision*, pages 462–469, 2005.

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

[5] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.

[6] H. Dee and D. Hogg. Detecting inexplicable behaviour. In *British Machine Vision Conference*, pages 477–486, 2004.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behaviour recognition via sparse spatio-temporal features. In *IEEE W. on Visual Surveillance*, pages 65–72, 2005.

[8] R. Donner, M. Reiter, G. Langs, P. Peloscheck, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, October 2006.

[9] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 838–845, 2005.

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

[11] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *IEEE International Conference on Computer Vision*, pages 742–749, 2003.

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

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

[14] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1031–1038, 2005.

[15] 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.

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

[17] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE Conference on Automatic Face & Gesture Recognition*, 2000.

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

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

[20] J. Kruskal and M. Liberman. *The symmetric time-warping problem: From continuous to discrete*. Addison-Wesley, 1983.

[21] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, June 2006.

[22] L.R.Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[23] 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.

[24] R. J. Morris and D. C. Hogg. Statistical models of object interaction. *International Journal of Computer Vision*, 37(2):209–215, 2000.

[25] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modelling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.

[26] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In K. Kurihara, editor, *Face Recognition*, pages 377–416. Advanced Robotics Systems, Vienna, 2007.

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

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

[29] Y. Shan, H. S. Sawhney, and A. Pope. Measuring the similarity of two image sequence. In *Asian Conference on Computer Vision*, 2004.

[30] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[31] J. Wilpon, L. Rabiner, C. Lee, and E. Goldman. Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Trans. Acoustic, Speech and Signal Processing*, pages 1870–1878, 1990.

[32] T. Xiang and S. Gong. Video behaviour profiling and abnormality detection without manual labelling. In *IEEE International Conference on Computer Vision*, pages 1238–1245, 2005.