How to Distinguish Posed from Spontaneous Smiles using Geometric Features

Michel F. Valstar
Department of Computing
Imperial College London
michel.valstar@imperial.ac.uk

Hatice Gunes
Department of Computer Systems
University of Technology, Sydney
haticeg@it.uts.edu.au

Maja Pantic
Department of Computing/EEMCS
Imperial College London, UK/Universiteit Twente, Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

Automatic distinction between posed and spontaneous expressions is an unsolved problem. Previously cognitive sciences' studies indicated that the automatic separation of posed from spontaneous expressions is possible using the face modality. However, little is known about the information from head and shoulder motion. In this work, we propose to (i) distinguish between posed and spontaneous smiles by fusing head, face, and shoulder modalities, (ii) investigate which modalities carry important information and how the modalities relate to each other, and (iii) to which extent the temporal dynamics of these signals attribute to solving the problem. A cylindrical head tracker is used to track head motion and two particle filtering techniques to track facial and shoulder motion. Classification is performed by kernel methods combined with ensemble learning techniques. We investigated two aspects of multimodal fusion: the level of abstraction (i.e., early, mid-level, and late fusion) and the fusion rule used (i.e., sum, product and weight criteria). Experimental results from 100 videos displaying posed smiles and 102 videos displaying spontaneous smiles are presented. Best results were obtained with late fusion of all modalities when 94.0% of the videos were classified correctly.

Categories and Subject Descriptors

I.2.10 [**Vision and scene understanding**]: [Video analysis]; H.1.2 [**User/Machine systems**]: [Human information processing, Human Factors]

General Terms

Human Factors, Algorithms, Experimentation

Keywords

Human information processing, Deception detection, Multimodal video processing

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan

1. INTRODUCTION

Human-to-human interaction is multimodal. People naturally communicate multimodally by means of language, tone, facial expression, gesture and head movement, body movement and posture and possess a refined mechanism for the fusion of these data. To date, machines are not well able to emulate this ability.

Psychological research findings suggest that humans rely on the combined visual channels of face and body more than any other channel when they make judgements about human communicative behaviour [2]. However, most of the existing expression analysers are monomodal and target human facial affect analysis, only learning to recognise a small set of prototypical emotional expressions [18]. To facilitate the detection of subtle facial signals like a frown or a wink instead of prototypical expressions several research groups have begun research on machine analysis of facial muscle actions. These atomic facial signals, or AUs, are defined in the Facial Action Coding System (FACS, [12]). Every possible facial expression can be described as a specific combination of AUs. A number of promising prototype systems have been proposed recently that can recognise 15 to 27 different AUs (from a total of 44 AUs) in either (near-) frontal view or profile view face image sequences [18].

In addition, independently of whether the approach is AU- or affect-oriented, most of the past work on automatic facial expression analysis is aimed at the analysis of posed (i.e., volitionally displayed) facial expression data. Only recently few works have been reported on machine analysis of spontaneous facial expression data (e.g. [3, 27]). However, to the best of our knowledge, no vision-based system exists yet that is based on FACS, takes multiple behavioural cues into account (e.g., facial, head an shoulder gestures), and automatically discerns between posed and spontaneous expressions.

Overall, computer vision and human-computer interaction (HCI) communities have not adequately exploited the expressive information carried by the body modality [15]. Attempts to recognise affective body movements are few and efforts are mostly on the analysis of posed body actions without considering the facial actions (e.g., [6]). Static postures of acted emotions were recorded by De Silva et al. [25] using a motion capture system, but the authors did not attempt to recognise the postures automatically.

Although it has been commonly stated that reliable assessment of human affect requires the concurrent use of mul-

tiple modalities [15], relatively few works have focused on implementing affect recognition systems using multimodal data [19]. Gunes and Piccardi [13] presented an approach to bimodal recognition of posed expressions of emotions by recording face and body gestures simultaneously using two cameras. Kapoor and Picard focused on the problem of detecting the level of interest in a child who is solving a puzzle [16]. They combined sensory information from the face video, the posture sensor (a chair sensor) and the game being played in a probabilistic framework. Karpouzis et al. [17] fused data from facial, bodily and vocal cues using a simple recurrent network to detect emotions. Cohn et al. [7] conducted a multimodal analysis of spontaneous smiles, comparing data from the face and the head modalities. However, in that study no fusion of the data was carried out. The major findings reported by these works were: when recognising affective states from multimodal nonverbal data, body gestures or postures provide better information than other modalities (i.e., face), and the fusion of multiple modalities significantly outperforms classification using the individual modalities [13, 16].

This paper reports on a method for automatic, multi-cue discrimination between posed and spontaneous smiles. We focus on smiles because of their importance in human development and communication. Developmentally, smiles are one of the first emotion expressions to appear, they occur with relatively high frequency throughout the lifespan and they express a multitude of meanings, including joy, appeasement and greetings, and they often serve to mask anger, disgust, and other negative emotions.

In our study, we explore the following three issues: Firstly, we want to know what the relative importance of the face, the head and the shoulders are for the problem of posed vs. spontaneous smile recognition. It is widely accepted that facial expressions reveal whether a display of affect is posed or genuine [8, 10, 14]. However, there is no such consensus when it comes to the relevance of bodily motion.

Darwin argued that because our bodily actions are easier to control on command than our facial actions, the information contained in the signal of body movements should be less significant than the face, at least when it comes to discerning spontaneous from posed behaviour [10]. Ekman however, argued that people do not bother to censor their body movements [10] and therefore, the body would be the more *'leaky'* source. Furthermore, research in nonverbal behaviour and communication theory stated that truthful and deceptive behaviour differ from each other in lack of head movement [5] and lack of illustrating gestures which accompany speech [9]. Therefore, we expect to find valuable information concerning the nature of a nonverbal expression (i.e., posed or spontaneous) in head and shoulder movements as well as in facial actions.

Secondly, we want to investigate the importance of the temporal dynamics of nonverbal behaviour for the problem of posed vs. spontaneous smile recognition. The body of research in cognitive sciences which suggests that the temporal dynamics of human facial behaviour are a critical factor for interpretation of the observed behaviour, is large and growing [1, 4, 14]. They are the key parameter in differentiation between posed and spontaneous facial expressions [10, 14]. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles, are slow in onset, can have multiple AU12 apices (multiple peaks in the mouth corner movement), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1s [8]. For the body modality, DePaulo et al. reported that deceivers' body actions appeared overcontrolled and abrupt [9]. Based on these findings, expect that the temporal dynamics will play a significant role in the recognition of posed vs. spontaneous smiles.

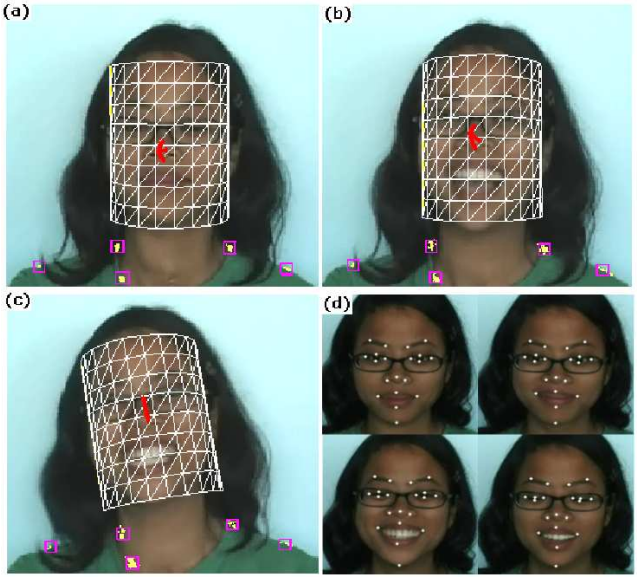

Figure 1: Illustration of the tracking procedure and the points used to obtain the tracking data: (a-c) for the head and shoulder modalities, and (d) for the face modality.

Thirdly, we will look into the effect that different multimodal data fusion strategies have on the classification accuracy of posed and spontaneous smiles and compare these with monomodal classification results. The fusion strategies differ from each other mainly in two aspects: the abstraction level of the used features and the way classification results are combined.

The remainder of this paper is organised as follows. Section 2 presents the various techniques employed for tracking the movement of the head, facial features, and shoulders. Section 3 describes how the recognition of the temporal phases (onset, apex offset) of these modalities is achieved. Section 4 discusses the strategies used for fusing the extracted data. Section 5 provides the experimental results and Section 6 presents the conclusions drawn from the study.

2. TRACKING

We employ a different tracker for each modality: a Cylindrical Head Tracker to track the head motion [29], Particle Filtering with Factorised Likelihoods to track 12 fiducial points in the face [20, 21], and Auxiliary Particle Filtering to track the shoulders motion [22]. Fig. 2 shows the facial and shoulder points that we track.

To capture the head motion we employ the Cylindrical Head Tracker developed by Xiao et al. [29]. The head tracker estimates the six degrees of freedom of head motion: horizontal and vertical position in the scene, distance to the

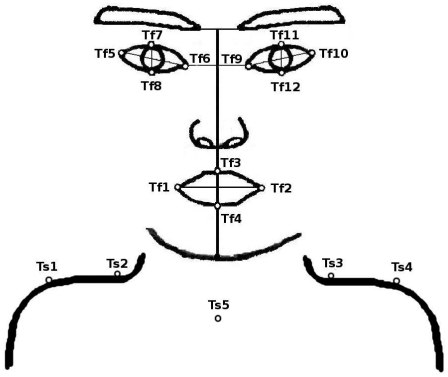

Figure 2: Tracked points $T_{f1} \dots T_{f12}$ of the face and tracked points $T_{s1} \dots T_{s5}$ of the shoulders.

camera (i.e. scale), pitch, yaw and roll. This is denoted as the set of parameters $T_h = \{T_{h1} \dots T_{h6}\}$ with dimensions $n * 6$. Here n is the number of frames of the input image. A cylindrical head model is manually fitted to the initial face region (see Fig. 1), and the face image is cropped and projected onto the cylinder as the template of head appearance. For any given subsequent frame, the face template is projected onto the image plane assuming that the pose has remained unchanged from the previous frame. Then, the difference between the projected image and the current frame is computed, providing the correction of the estimated head pose.

To capture all facial motion that is characteristic for smiles, we track four points on the mouth and eight points in the eye region (see Fig. 2). For the mouth we track the mouth corners, the upper, and the lower lip. For the eyes we tracked the inner and outer eye corners as well as the upper and lower eyelids. The algorithm we used to track these facial points is Particle Filtering with Factorised Likelihoods (PFFL) [20]. The facial points are located in the first frame using an automatic facial point detector [28]. The size of all the faces are scaled using the interocular distance (IOD), defined as the distance between the inner eye points. We used the head tracker data to register the images so that the face has the same pose in every frame. We subsequently track each colour template in the rest of the image sequence with the PFFL algorithm. We used the observation model proposed in [21], which is both insensitive to variations in lighting and able to cope with small deformations in the template. This polymorphic aspect is necessary as many areas around facial points change their appearance when a facial action occurs (e.g. the mouth corner in a smile). The facial point tracking scheme results for every image sequence in a set of points $T_f = \{T_{f1} \dots T_{f12}\}$ with dimensions $n * 12 * 2$.

The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 2). The stable point is used to remove any rigid motion of the torso (see section 4). We use standard Auxiliary Particle Filtering (APF) [22] instead of PFFL because it is less complex and faster than PFFL, it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. The shoulder tracker results in a set of points $T_s = \{T_{s1} \dots T_{s5}\}$ with dimensions of $n * 5 * 2$.

3. TEMPORAL SEGMENTATION

A face or body action can be in any one of four possible phases: (i) the onset phase, where muscles are contracting and the changes in appearance are growing stronger, (ii) the apex phase, where the face/body action is at a peak and there are no more changes in appearance due to this particular action, (iii) the offset phase, where the muscles responsible for the face/body action are relaxing and the face/body returns to its original, neutral appearance, and (iv) the neutral phase, where there are no signs of activation of the investigated face/body action.

According to [8, 10, 27], timing, duration and speed of facial actions are highly important cues for distinguishing posed from spontaneous facial expressions. Especially the speed at which a facial expression develops or diminishes during respectively the onset and offset phases has proved to be highly discriminative. In section 4 we will define these cues in terms of the speed of tracked points or, in the case of the head, the angular and translational velocity of the entire head. In order to be able to compute these cues, or attributes, separately for the onset, apex and offset temporal segments of a face/body action, we first need to know when every temporal segment of every modality begins and ends.

For the recognition of the temporal segments of the face, we adopted the method proposed in [26]. This method computes for every frame of an input face video a set of spatio-temporal attributes based on the tracked facial points, which are then passed on as input to a multiclass classifier that decides what temporal segment of what AU the frame belongs to. The multiclass classifier used is a one-vs-one Gentle Support Vector Machine (GentleSVM). We detect temporal segments for AU6, AU12 and AU13. This segmentation results in a number of $m_f = m_{AU6} + m_{AU12} + m_{AU13}$ temporal segments for the face modality.

The temporal segments of the head and the shoulder modalities are obtained using a rule-based expert system as we do not have the manually labelled temporal segment data for these modalities based on which an automatic detector could be trained. The temporal segments are found as follows. We only consider one action to be possible for the head and the shoulders, that is we only check if the head or the shoulders are in their neutral position or not. Given the displacement of an arbitrary time series z of vectors with length n:

$$\delta(z, t) = z(t) - z(1) \tag{1}$$

we find:

$$q_{b1}(t) = \|\delta(b_1, t)\| \tag{2}$$
$$r_{b1}(t) = \|d(b_1, t)\|/dt \tag{3}$$
$$q_{b2}(t) = \|\delta(b_2, t)\| \tag{4}$$
$$r_{b2}(t) = \|d(b_2, t)\|/dt \tag{5}$$

where $b_1 = \{T_{h4}, T_{h5}, T_{h6}\}$ is the time series of pitch, roll and yaw and $b_2 = \{T_{h1}, T_{h2}, T_{h3}\}$ is the vector of the head positions. Both are subsets of T_h. Now, for each time t we say that the head is in its neutral phase if both the angle difference and the head translation are close to zero, that is, IFF $q_{b1} < \theta_1$ AND $q_{b2} < \theta_2$. If the head is not in its neutral phase, we continue to check whether the head is in its apex phase, that is, whether the angular velocity or the

translational velocity are sufficiently close to zero. This is the case IFF $r_{b1} < \theta_3$ AND $r_{b2} < \theta_4$. If the head is neither in its neutral nor in its apex phase, we check the sign of the first derivative to time of the angular motion $d(\delta(r_{b1}, t))/dt$ and the current position of the head. If the head is moving away from it's neutral position, we assign the onset phase, otherwise we assign the offset phase. This results in m_h temporal segments for the head modality.

To find the temporal segments for the shoulders, we use the same strategy as for the head. Now, $h_1(t)$ denotes the angle made by the horizontal axis and the line connecting shoulder points T_{s1} and T_{s2} at time t and $h_2(t)$ is the angle made by the horizontal axis and the line connecting shoulder points T_{s3} and T_{s4} at time t. This results in m_s temporal segments for the shoulder modality, and in total $m = m_f + m_h + m_s$ temporal segments for all modalities together. The values of the thresholds $\delta_1 \dots \delta_4$ were found using cross validation during training.

4. FUSION STRATEGIES

As outlined in the introduction, one of the goals of this paper is to investigate the effects on the recognition performance by using different abstraction levels of the feature definition, different abstraction levels of the classification schemes, and different fusion rules. In order to achieve this, we implement three different fusion strategies to tackle the problem of multimodal posed vs. spontaneous smile recognition; early, mid-level and late fusion. We distinguish two types of features. Early fusion uses low-abstraction features while mid- and late level fusion use high-abstraction features.

4.1 Early fusion

In early fusion, the elementary attributes of each modality are combined into one low-abstraction feature vector, which serves as the input to one classifier. In our case, the elementary attributes are simple operations on the tracking data, such as the distances between two tracked points or the angular velocity of the pitch of the head. Features are computed and passed to the classifier on a per-frame basis, resulting in a time-series of classification predictions $\mathbf{y}$ with length n for every input video.

From the head modality we simply used the output from the tracker (see section 2), $f_h = T_h$.

From the face modality, we concatenated the x- and y-values of all tracked points, the distances between all pairs of points and the angles between the line connecting two points and the horizontal axis:

$$\begin{aligned} f_f(t) = \{&T_{f1,x}(t), T_{f1,y}(t), \dots, T_{f12,y}(t), \\ &\eta(T_{f1}(t), T_{f2}(t)), \dots, \eta(T_{f11}(t), T_{f12}(t)), \\ &\alpha(T_{f1}(t), T_{f2}(t)), \dots, \alpha(T_{f11}(t), T_{f12}(t))\} \end{aligned} \quad (6)$$

where $T_{fi,d}(t)$ is the value of tracked point i at time t for dimension d (either x or y), $\alpha(p_1, p_2)$ denotes the angle defined by the line connecting two points and the horizontal axis, and $\eta(p_1, p_2)$ denotes the Euclidean distance between two points. For the shoulder modality, we defined the feature vector as follows:

$$\begin{aligned} f_s(t) = \{&\alpha(T_{s1}(t), T_{s2}(t)), \alpha(T_{s4}(t), T_{s3}(t), \\ &[\delta(T_{s1,y}, t) + \delta(T_{s2,y}, t)], [\delta(T_{s3,y}, t) + \delta(T_{s4,y}, t)]\} \end{aligned} \quad (7)$$

See Fig. 2 for the numbering of the shoulder points. To obtain our final feature vector used in early fusion, we first concatenated the above attributes as $f = \{f_h, f_f, f_s\}$. Then, we defined

$$F_e(t) = \{f(t), d(f(t))/dt, \delta(f, t)\} \quad (8)$$

In this definition of our features, we denominated the first term of the right hand side of eq.(8) as the static features and the second and third terms as the dynamic features.

The feature vector $F_e(t)$ serves as input to a GentleSVM-Sigmoid classifier. This tandem arrangement of feature selection using GentleBoost and classification using SVMs has been shown to attain high classification rates [26]. Unfortunately the output of an SVM is not a good measure for the posterior probability of its prediction. Therefore we pass the output of the SVM to a sigmoid function that has been shown to provide a reasonable measure for the posterior probability [23]. We will refer to this feature selection-classifier combination as a GentleSVM-Sigmoid classifier. The vector $F_e(t)$ provides the t^{th} element of the time series of predictions $\mathbf{y}$. Under a Maximum-a-Posteriori (MAP) approach, $\mathbf{y}(t)$ must be assigned to one of the two possible classes (posed or spontaneous), according to maximum posterior probability.

4.2 Mid-level fusion

Mid-level fusion attains a higher level of data abstraction within every modality, yet we still fuse all attributes into one vector, and use only one classifier. Often the elementary attributes of early fusion are used to define a set of abstract symbols (such as AUs), or they are used to compute more heuristic features. In our approach, we transform the elementary attributes derived previously into both symbols and higher level features. The symbols we derive at this stage are the temporal segments of AU6, AU12 and AU13 and the temporal segments of the head and shoulder action. We refer to the combination of symbols and higher level features as the high-abstraction features.

For each temporal segment, we define attributes based on the works of Cohn and Schmidt, Ekman, and Valstar et al. [8, 10, 27]. These include the morphology, speed, symmetry, duration, apex overlap (i.e., number of frames that two actions are in apex simultaneously), trajectory of a face/body action and the order in which the temporal segment of the different face/body actions occur. Similar to the case of early fusion, we concatenate the attributes of all modalities into one vector, which serves as the input to a GentleSVM-Sigmoid classifier.

Because the GentleSVM-Sigmoid classifier requires the input vectors to be of the same length, we were forced to change our definition of the temporal segments. Face/body actions frequently have multiple apices before returning to its neutral position. We decided to force any sequence of temporal segments into the temporal pattern onset-apex-offset. To do so, we chose the apex segment with the longest duration as the apex phase of our new forced temporal pattern. The beginning of the new — forced — onset is chosen as the first non-neutral frame before the apex. Conversely, the end of the new forced offset is chosen as the last non-neutral frame after the apex phase. The neutral segments are discarded as, by definition, they should not contain any

information. This results in k forced temporal segments for all modalities.

When enforcing this particular temporal pattern, information about the original temporal pattern of the symbols is lost. In order to retain some of this information, we compute a number of concurrency attributes M for every video. The concurrency feature vector M is computed before we enforce the new temporal pattern and contains the duration of all symbols, the order in which the symbols are activated and the duration of overlap of the symbols.

We thus define for every temporal segment of the forced pattern the following set of features. For each segment of a face action (either AU6, AU12, or AU13) or shoulder action (motion) we define the mean/max displacement and the mean/max velocity of the tracked properties during that segment as:

$$g_{symbol,segment} = \left\{ \frac{\sum_{t=t_1}^{t_m} \delta(T,t)}{t_m + 1 - t_1}, \max_{t=t_1 \ldots t_m} \delta(T,t), \frac{\sum_{t=t_1}^{t_m} d(T(t))/dt}{t_m 2 + 1 - t_1}, \max_{t=t_1 \ldots t_m} d(T(t))/dt \right\} \quad (9)$$

where t_1 and t_m are the first and the last frames of a temporal segment, respectively. T is the tracking data of the eye points when considering AU6, the mouth points when considering AU12 or AU13, the head tracking when considering head action, and the shoulder tracking data when considering shoulder action. Additionally, for the face and shoulder modalities we compute the asymmetry value a. For the head modality, we define a to be an empty set. The final feature vector for mid-level fusion is thus found as the union of M, g and a for all symbols S ={AU6,AU12,AU13, head action, shoulder action} and for all forced temporal segments R ={onset, apex, offset}:

$$F_m = \left\{ \bigcup_{i \in S, j \in R} (g_{i,j}), \bigcup_{i \in S, j \in R} (a_{i,j}), M \right\} \quad (10)$$

Because the values of all features depend on the time parameters t_1 and t_m we consider all mid-level parameters to be temporal dynamic.

Given the feature vector F_m, where F_m describes an entire smile, we again use a GentleSVM-Sigmoid classifier to predict the class of the video under the previously described MAP approach.

4.3 Late fusion

Late fusion is similar to mid-level fusion in the sense that we again attain a high level of data abstraction within every modality. However, in the case of late fusion we also attain a higher level of abstraction for the classification procedure, computing a separate posterior probability for each temporal segment. This removes the need to enforce the strict onset-apex-offset temporal pattern used in mid-level fusion. Indeed, we will use all m temporal segments, collected from all modalities. In this way we obtain a variable number of predictions y for every video. This enables the system to discard a modality when needed (e.g., when the shoulders move out of view), as the fusion rules we employ are invariant to the number of inputs. We compute a concurrency feature vector M that contains the duration of all temporal segments each modality, the order in which the segments are activated and the overlap of the apex phases of every combination of symbols.

Table 1: Description of the three late fusion criteria used: sum, product and weight.

sum	$k = \text{argmax}_{k=1}^{2} p(w_k \mid F_{fl} + F_{hl} + F_{sl})$
product	$k = \text{argmax}_{k=1}^{2} p(w_k \mid F_{fl} * F_{hl} * F_{sl})$
weight	$k = \text{argmax}_{k=1}^{2} \sigma_f p(w_k \mid F_f)$ $+\sigma_h p(w_k \mid F_{hl}) + \sigma_s p(w_k \mid F_{sl})$

Every temporal segment from every symbol generates one feature vector. This vector is defined as:

$$F_l(i) = \{g(i), a(i)\} \quad (11)$$

where i is a temporal segment. Again, all features are considered dynamic. Each feature vector $F_l(i)$ is used as input to the appropriate GentleSVM-Sigmoid classifier. That is, we train a different classifier for every temporal segment type for every symbol. Thus we train one GentleSVM-Sigmoid for the onset phase of AU12, one for the apex phase of shoulder actions, etc. A separate classifier is trained for the concurrency attributes. The vector $F_l(i)$ then provides the i^{th} element of predictions y. After achieving this, the general approach of late fusion of the individual classifier outputs can be described as follows.

The time series y represents the whole image sequence and $F_l = (F_{fl}, F_{hl}, F_{sl})$ represent the overall feature vectors consisting of the face F_{fl}, head F_{hl}, and shoulder F_{sl} feature vectors. Under a Maximum-a-Posteriori (MAP) approach, y must be assigned to one of the two classes (w_1, w_2), having maximum posterior probability $p(w_k|y)$. Once the posterior probabilities per modality per temporal segment are obtained by again passing the output of the SVM to a sigmoid function as proposed by Platt [23], late fusion is applied. The three separate classifiers provide the posterior probabilities $p(w_k|F_{hl})$, $p(w_k|F_{fl})$ and $p(w_k|F_{sl})$ for the head, face and shoulder modalities, respectively, to be combined into a single posterior probability $p(w_k|y)$ with one of the fusion methods described in Table 1. Note that the weights are derived from the classification results on the training data during cross validation.

5. RESULTS

We evaluated the three fusion approaches on 100 videos of posed smiles and 102 videos of spontaneous smiles using 10-fold cross-validation. The videos were taken from the MMI-facial expression database. All videos were recorded from a near-frontal view, under controlled lighting conditions. During the recordings, the subjects for the posed dataset were asked to show a series of facial actions, one of which was a smile. The subjects in the spontaneous videos were recorded while watching cartoons or clips of nauseating footage for about 10 minutes. All videos were edited to ensure that they contained exactly one smile. Multiple apices were allowed, the video was only cut when the face had returned to its neutral phase.

Table 3 shows the classification results for all fusion strategies. To our best knowledge, the system presented here is the first to propose discerning posed from spontaneous smiles by

Table 2: Selected low-abstraction features to distinguish posed from spontaneous smiles.

Relevance	Modality	Feature definition
1	Face:	$\delta(\eta(T_{f3}(t), T_{f4}(t)), t)$
2	Shoulders:	$T_{f2,y}(t)$
3	Face:	$\delta(\alpha(T_{f1}(t), T_{f4}(t)), t)$
4	Head:	$T_{f3,x}(t)$
5	Face:	$\delta(T_{f2,x}(t), t)$

Table 3: Classification, recall and precision rates for the different fusion strategies employed.

Fusion strategy	Cl. rate	Recall	Precision
Early	0.886	0.889	0.880
Mid-level	0.881	0.883	0.886
Late (sum)	0.931	0.956	0.920
Late (product)	0.940	0.964	0.933
Late (weight)	0.931	0.943	0.927

fusing video data from face, head and body actions. All results were obtained using 10-fold cross-validation. For the purpose of computing the precision and recall, we considered spontaneous smiles to be the positive class. Overall we can say that the proposed system works as desired, being able to discern between posed and spontaneous smiles with fairly high accuracy. There is no significant difference between early and mid-level fusion at a 5% significance level. Late fusion does score significantly higher than early fusion and mid-level fusion.

The high results for late fusion could be explained by two factors. First there is the high classification abstraction. Specialised classifiers are learnt to distinguish posed from spontaneous smiles for each segment of a smile, i.e., during the onset of head motion, the apex of a smile, etc. Because all specialised classifiers return a posterior probability, the fusion rule can then be used to generalise from the results per segment of an action to the entire action (i.e. a smiling face with its accompanying bodily action).

The second explanation for the high score for late fusion is the high data abstraction. The low-abstraction features only describe simple attributes: positions, distances and angles of points. Moreover, they only describe those attributes at one point in time — the frame for which they are defined. The high-abstraction features capture more general physical phenomena such as the duration of a temporal segment, the average speed during onset and the order in which actions occur.

To investigate what the relative importance of each modality was, we adapted the early fusion strategy. We performed seven tests, each time using a different combination of modalities. The results of this test are shown in Table 4. For enhanced resolution, the results listed are computed per frame basis, instead of per video. In addition, Table 5 provides a matrix showing which of the results were statistically different on a 5% significance level. When we regard only single modalities (combinations I, II and III), the head modality performs best according to our results. However, the recognition rates between the separate modalities are not significantly different at a 5% significance level ($P = 0.05$). Early fusion of all modalities (combination VII) *is* significantly better then any of the single-modality combinations.

Table 4: Comparison of classification rate, recall and precision of the different modalities separately and fused.

Modality	Cl. rate	Recall	Precision
I Face	0.812	0.841	0.868
II Head	0.822	0.823	0.916
III Shoulders	0.794	0.793	0.915
IV Face-Head	0.867	0.897	0.893
V Face-Shoulders	0.871	0.896	0.899
VI Head-Shoulders	0.845	0.861	0.899
VII All	0.895	0.919	0.916

Table 5: Matrix of statistical significant different classification rates. Roman indices relate to the modality combinations listed in table 4. A 1 indicates statistically significantly different results.

	I	II	III	IV	V	VI	VII
I	0	0	0	0	1	0	1
II	0	0	0	0	1	0	1
III	0	0	0	1	1	1	1
IV	0	0	1	0	0	0	0
V	1	1	1	0	0	0	0
VI	0	0	1	0	0	0	1
VII	1	1	1	0	0	1	0

To further investigate the relevance of the different modalities, we performed an analysis of the features selected by GentleBoost. For early fusion, 62% of the selected features originated from the face modality, 16% from the head modality, and 22% origined from the shoulder modality. For mid-level fusion, 40% of the features came from the face modality, 40% from the head modality, 13,3% from the shoulder modality and 6,7% of the selected features origined from the concurrency features (which span all modalities). For early fusion GentleBoost selected 45 low-abstraction features, of which the first 5 are listed in Table 2. The first 5 selected features for mid-level fusion are listed in Table 7.

In late fusion, feature selection takes place in the separate classifiers specialised for each temporal phase of each modality so a comparison of the selected features is not feasible. However, we might learn something from the classification performance of the specialised classifiers. Table 6 lists the classification results attained when we use only one specialised classifier to classify an entire video into a posed or a spontaneous smile. From this table, we can read two things: first, the head modality seems to be most reliable in late fusion. Second, the offset phase seems to carry the least information.

Table 6: Classification rates for the specialised classifiers used in late fusion. There is no temporal phase associated with the concurrency classifier.

	Onset	Apex	Offset
Face	0.719	0.612	0.451
Head	0.781	0.826	0.742
Shoulders	0.752	0.766	0.638
Concurrency	0.781		

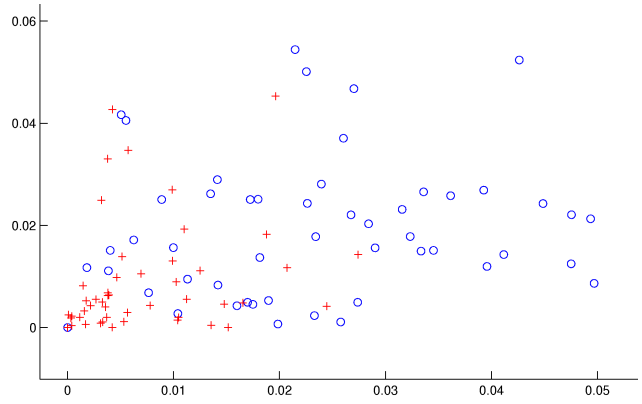

Figure 3: Mean velocity of the right mouth corner in x-direction during onset (x-axis) vs. mean velocity of the right mouth corner in x-direction during offset (y-axis). Crosses denote spontaneous smiles.

Based on the results for feature selection and the results of combinations of modalities, given our data, we might conclude that the head is the most important modality for distinguishing between posed and spontaneous smiles. This is in agreement with both other HCI works [13, 16] and cognitive scientists' works [5, 11]. But more importantly, the results show that the modalities complement each other. The result of all modalities combined (the fused result, same as the early-fusion result of table 3) is significantly better at $P = 0.05$ than any of the other modalities separately. This confirms our hypothesis that a multimodal approach benefits posed vs. spontaneous smile detection.

To answer our question regarding the relevance of temporal dynamics for automatic multimodal posed vs. spontaneous smile recognition, we again provide an analysis of the feature selection process. Table 2 shows all the selected features for low-abstraction features and Table 7 those for high-abstraction features, including the modality that the feature originated from. In the case of high-abstraction features, the originating temporal segment is listed as well. For early fusion, 24.4% of the selected features were static features, while 75.6% of the features were dynamic features.

While the fraction of static features was greater then we expected, we can still clearly see that the temporal dynamics are the most important features for automatic multimodal posed vs. spontaneous smile recognition. This is also reflected in the high classification results of late fusion, which uses only temporal dynamics. Fig. 3 shows a scatter plot of the mean velocity of the right mouth corner in x-direction during onset and offset. As we can see, spontaneous smiles have both a slower onset and a slower offset, consistent with the cognitive sciences' findings [8].

Ekman predicted that the asymmetry of facial actions is an indicator for distinguishing posed from spontaneous expressions [10]. We did not find any evidence for this however, only one of the selected high-level features was an asymmetry feature. Although this observation is in disagreement with Ekman's findings, the same lack of correlation between asymmetry and the nature of the expression was previously reported by Schmidt et al. [24].

Table 7: Selected high-abstraction features to distinguish posed from spontaneous smiles.

Rel.	Mod./seg.	Feature
1	onset head:	mean angular velocity
2	concurrency:	order apex shoulders
3	apex head:	max translational displ.
4	apex head:	mean angular velocity
5	apex shoulders:	max angular velocity of left shoulder

6. CONCLUSION

We have shown that our proposed multimodal approach to automatic distinction between posed and spontaneous smiles is extremely accurate. From the results presented, it is clear that fusing video data from the face, head and shoulders increases the accuracy. This is in agreement with the body of work in cognitive sciences indicating that humans leak their intentions not only through facial expressions, but also through their body language. It is hard to say which modality is the most important. The results seem to indicate that the head is the most reliable source, followed closely by the face. However, more experiments are needed to confirm this. Dynamic attributes are clearly more important than static ones. This can be seen from the large number of dynamic features selected during early fusion, as well as from the high results for late fusion. Regarding the different fusion strategies, late fusion clearly performs best. The first reason for this is that with late fusion we are able to decompose the problem in smaller subproblems, for which we can train specialised classifiers. Another major benefit of late fusion is the use of high-abstraction features, which encode important temporal dynamic attributes of human nonverbal behaviour.

7. REFERENCES

[1] Z. Ambadar, J. Schooler, and J. F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expessions. *Psychological Science*, 16(5):403–410, 2005.

[2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta–analysis. *Psychological Bulletin*, 11(2):256–274, 1992.

[3] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[4] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *J. Experimental Psychology*, 4(3):373–379, 1978.

[5] D. Buller, J. Burgoon, C. White, and A. Ebesu. Interpersonal deception: Vii. behavioral profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology*, 13(5):366–395, 1994.

[6] A. Camurri, G. Volpe, G. D. Poli, and M. Leman. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1):43–53, 2005.

[7] J. Cohn, L. Reed, T. Moriyama, J. Xiao, K. Schmidt, and Z. Ambadar. Multimodal coordination of facial action, head rotation, and eye motion during

spontaneous smiles. In *Proc. of the Sixth IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FG'04)*, pages 129 – 138, 2004.

[8] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *J. Wavelets, Multi-resolution and Information Processing*, 2(2):121–132, 2004.

[9] B. DePaulo. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003.

[10] P. Ekman. Darwin, deception, and facial expression. *Annals of New York Ac. of sciences*, 1000:105–221, 2003.

[11] P. Ekman and W. V. Friesen. The repertoire of of nonverbal behavior: categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.

[12] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System*. A Human Face, 2002. Salt Lake City.

[13] H. Gunes and M. Piccardi. Bi–modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 2006. In Press,doi:10.1016/j.jnca.2006.09.007.

[14] U. Hess and R. E. Kleck. Differentiating emotion elicited and deliberate emotional facial expressions. *European J. of Social Psychology*, 20(5):369–385, 1990.

[15] E. Hudlicka. To feel or not to feel: the role of affect in human–computer interaction. *Int. Journal of Human–Computer Studies*, 59(1–2):1–32, 2003.

[16] A. Kapoor and R. Picard. Multimodal affect recognition in learning environments. In *Proc. of the ACM Int'l Conf. on Multimedia*, pages 677–682, 2005.

[17] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal and bodily expressions recognition. *Lecture Notes in Artificial Intelligence*, 4451:92–116, 2007.

[18] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Kurihara, editor, *Face Recognition*, pages 327–366. ARS Press, 2007. Vienna, Austria.

[19] M. Pantic, N. Sebe, J. Cohn, and T. Huang. Affective multimodal human–computer interaction. In *Proc. of the ACM Int'l Conf. on Multimedia*, pages 669–676, 2005.

[20] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. *Proc. Int'l Conf. Automatic Face & Gesture Recognition*, pages 97–102, 2004.

[21] I. Patras and M. Pantic. Tracking deformable motion. *Proc. Int'l Conf. Systems, Man and Cybernetics*, pages 1066–1071, 2005.

[22] M. K. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filters. *J. Am. Statistical Association*, 94(446):590–616, 1999.

[23] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT press, 2000. Cambridge, MA.

[24] K. Schmidt, Z. Ambadar, J. Cohn, and I. Reed. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *J. Nonverbal Behavior*, 30(1):37–52, 2006.

[25] P. R. D. Silva, A. Kleinsmith, and N. Bianchi-Berthouze. Towards unsupervised detection of affective body posture nuances. *Int. Conf. Affective Computing and Intelligent Interaction*, pages 32–40, 2005.

[26] M. F. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, page 149, 2006.

[27] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. *Proc. ACM Intl. conf. on Multimodal Interfaces*, pages 162–170, 2006.

[28] D. Vukdadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted features. *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, pages 1692–1698, 2005.

[29] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int'l. J. Imaging Systems and Technology*, 13(1):85–94, 2003.