The influence of temporal facial information on the classification of posed and spontaneous enjoyment smiles

Miriam W. Huijser
10167218

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligentie

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Prof. dr. T. Gevers

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

June 27th, 2014

Abstract

It is important to train classifiers that are able to make a distinction between posed and spontaneous facial expressions for the improvement of facial expression classification In this thesis, the focus is on the classification of posed and spontaneous enjoyment smiles using temporal facial information. Four experiments are conducted to investigate the influence of temporal information from the head, lips and eyes, and the influence of the fusion of these sources of information, on the classification of posed and spontaneous enjoyment smiles. Continuous Hidden Markov models (CHMM) and Support Vector Machines (SVM) are trained on features extracted from smile videos from the UvA-NEMO Smile Database (Dibeklioğlu et al., 2012). In the conducted experiments, features from the lips are found to be the most discriminating features for classification, followed by features from the head and from the eyes. Furthermore, the fusion of sources of information generally leads to a better classification than each of the features on their own. Despite the low performance of the CHMM classifiers which achieved a maximum recognition rate of 54%, the SVM classifiers achieved a highest recognition rate of 78%. When compared to posed and spontaneous smile classification in still images, it is shown that the use of temporal information improves the classification of posed and spontaneous enjoyment smiles with a recognition rate improvement of 4%.

Contents

Acknowledgement

I would like to thank my supervisor Prof dr Theo Gevers for his support and guidance. Likewise, I would like to thank Hamdi Dibeklioğlu for providing me with the MATLAB-code to normalize the facial landmarks and for answering any of my questions regarding the facial landmarks corresponding to the faces in the UvA-NEMO Smile Database. In addition, I would like to thank Sezer Karaoglu for giving me a crash course in the toolbox used to train the SVM classifiers. Finally, I would like to thank my family and my boyfriend for motivating and supporting me, and my boyfriend in particular for employing his graphical talents to help me make the pictures for the eye features.

1 Introduction

The human face is an important source of information about gender, age, ethnicity and identity. In addition, it can reveal a great deal about a person's mood or emotions by means of facial expressions. Facial expressions are a form of nonverbal communication and play a vital role in social interaction between humans. Without the ability to show or identify facial expressions humans would be far less capable of expressing themselves, understanding others and having interpersonal relationships.

For half a century, researchers in the field of Artificial Intelligence primarily focused on providing computers logical-mathematical and linguistic intelligence and, conversely, the lack of emotional intelligence causing an unnatural interaction between humans and computers (Picard, 2004). Since the 1990s, however, a great deal of research on identifying facial expressions has been conducted for the improvement of human-computer interaction (HCI) (Lien, 1998; Shan et al., 2006; Picard, 2004). To achieve a more natural interaction between humans and computers, computers are trained to identify a person's feelings by classifying his/her facial expressions. This ability to sense human emotions by classifying facial expressions is important for helping computers to choose appropriate behaviour, for example, by being more helpful and behaving in a way that causes less aggravation in situations of distress (Picard, 2004).

The problem with this approach, however, is that facial expressions do not always correlate with the emotional feeling being experienced, as expressions can be posed rather than being spontaneous[1]. Hence, the classification of facial expressions is only useful for human-computer interaction when a distinction can be made between posed and spontaneous facial expressions.

One of the most frequently studied facial expressions for the analysis of spontaneity is the smile, because it is the most frequently used facial expression and the easiest facial expression to perform deliberately (Ekman, 2009). For these reasons, the focus of this study will be on the smile, more specifically, on smiles of enjoyment. These are smiles that are associated with feelings of happiness and, thus, differ from smiles of embarrassment or smiles that are used to mask sadness or anger (Ekman and Friesen, 1982; Keltner, 1995).

Existing research into smile spontaneity has either a psychophysical or a computational foundation (Ekman and Friesen, 1982; Krumhuber and Manstead, 2009; Valstar et al., 2007; Dibeklioğlu et al., 2012). The psychophysical research focuses on the perceived differences between posed and spontaneous smiles. In contrast, computational research focuses on algorithms and on extracting certain features that are often suggested as being discriminating by psychophysical research. These features are then used to train these algorithms to automatically distinguish between posed and spontaneous smiles.

In a great deal of psychophysical research on smile spontaneity, participants are instructed to pose a smile or their smile is invoked by showing them amusing material (Krumhuber and Manstead, 2009). Subsequently, researchers attempt to discover significant differences between posed and spontaneous smiles in the recorded smiles. Furthermore, in some studies participants are instructed to rate spontaneous and posed smiles presented to them (Krumhuber and Kappas, 2005; Krumhuber and Manstead, 2009; Del Giudice and Colle, 2007). These faces may be partially visible, for example, faces of which merely the mouth or the eyes are shown. The participants' ratings are then analyzed in an attempt to discover discriminating features of smiles used by people to distinguish between posed and spontaneous smiles.

Initially, classifiers of enjoyment smile classification are trained on features extracted from static faces (still images). The disadvantage of these classifiers is that they do not generalize well to dynamic faces in videos that display the development and fading of the smile because they are trained on static

[1] N.B. although facial expressions can be classified as genuine or insincere instead of spontaneous or posed, I have chosen the latter terms because genuineness analysis is less objective than spontaneity analysis. Whereas it is more complex, and perhaps even impossible, to discover whether emotions perceived on a person's face are truly present or absent, it is less problematic to discover whether a perceived facial expression has emerged spontaneously or not.

faces that merely display one pose in the entire process of smiling. Furthermore, because these classifiers are trained on static faces, the facial dynamic information provided during a smile is lost and it is this information that may be useful for improving classification.

More recently, the focus of computational research on enjoyment smile classification has shifted to classification using dynamic features from facial expression videos. These classifiers make use of dynamic facial features such as the speed and timing of the contraction of certain facial muscles (Cohn and Schmidt, 2004; Valstar et al., 2007; Dibeklioğlu et al., 2012). The disadvantage of these classifiers, however, is that dynamic temporal information of the face is often only partially represented. This is because the smile can be divided into three phases: the onset, apex and offset. Subsequently, features are computed once for each of the three phases, or the mean feature of each phase is calculated and used to represent the phase. This approach therefore ignores a great amount of information per smile phase and, therefore, per smile.

In the present study, a more continuous approach to temporal modelling is adopted in order that temporal facial information is more explicitly preserved for improved classification of posed and spontaneous smiles. Consequently, the principal question of the present study is whether the use of temporal facial information improves the classification of posed and spontaneous enjoyment smiles.

The method comprises data registration, temporal segmentation of the smiles, feature extraction, training of the Continuous Hidden Markov models and Support Vector Machines, and the (separate) fusion of these models trained on different features. Fusion of continuous Hidden Markov Models can either be done by fusing the maximum likelihood estimates or by majority voting using the classification result of each model. In contrast, fusion for Support Vector Machines basically means training on the concatenation of different features (see Section 3 for a detailed description of the method). The design criteria of the method are invariance to scale, meaning that a fixed distance between the head and the camera is not required, invariance to head movements, robustness against illumination and free smile duration.

1.1 Outline

In Section 2, firstly an overview is given of features analyzed in the literature and, subsequently, the most promising dynamic features of which the influence on classification will be investigated are listed. In Section 3, an overview of the method is given and discussed in more detail. In Section 4, the experiments are described, including the employed dataset, the metrics and the results. The results are further discussed in Section 5. Finally, the conclusion of this study is presented in Section 6 and directions for future research are presented in Section 7.

2 Related work

The facial movements of a smile and other facial expressions can be encoded by the Facial Action Coding System (FACS), which was originally introduced by Hjortsjö (1969) and further developed by Ekman and Friesen (1978). Each basic facial movement (action) caused by the contraction of individual facial muscles or a group of facial muscles is assigned an Action Unit (AU). In general, a smile can be identified by the contraction of the *zygomatic major* muscle, which raises the mouth corners, and this corresponds to Action Unit 12 in the FACS. As both spontaneous and posed smiles involve Action Unit 12, several other characteristics of spontaneous and posed smiles have been analyzed in the literature and experimented with in order to distinguish between the smiles.

Numerous psychophysical studies claim that the D-marker is a proper indicator of true enjoyment smiles. They are called Duchenne smiles in honour of Guillaume Duchenne, who laid the foundation of research on smile genuineness. The D-marker corresponds to Action Unit 6, which is the contraction of the *orbicularis oculi, pars lateralis* muscle that raises the cheek, narrows the opening of the eyes, and forms wrinkles (crow's feet) around the eyes (see Figure 1). According to Ekman and Friesen (1982) spontaneous enjoyment smiles indeed involve both the zygomatic major as well as the D-marker, whereas posed smiles involve the zygomatic major but lack the D-marker. However, in experiments of Krumhuber and Manstead (2009) the D-marker is about as frequently present in spontaneous as in posed smiles. In addition, the study by Schmidt et al. (2009) finds that although 96 percent of the participants shows the D-marker in their spontaneous smiles, 56 percent of the participants also show the D-marker in their deliberate smiles.

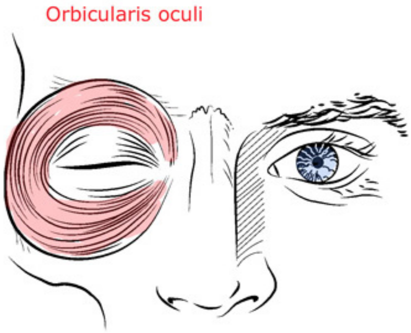

Figure 1: The orbicularis oculi muscle. Retrieved from `http://antranik.org/muscles-of-the-head/`

Another feature of smiles analyzed in the literature is smile symmetry. According to Ekman and Friesen (1982) spontaneous smiles are more symmetrical than posed ones. However, in later studies by Schmidt et al. (2009); Dibeklioğlu et al. (2012) a significant difference in symmetry between posed and spontaneous smiles is not observed. Hence, smile symmetry is not used to discriminate between posed and spontaneous smiles in this study.

Furthermore, the present study will focus less on static characteristics such as smile symmetry and more on dynamic characteristics of smiles. Examples of dynamic smile characteristics analyzed in the literature are the duration of the smile and the relative durations of the three phases of the smile: the onset, apex (peak) and offset phases. According to Ekman and Friesen (1982) felt smiles have a duration of two-thirds of a second to four seconds, as opposed to posed smiles that have a duration outside of this interval. This claim is in accordance with findings from Krumhuber and Manstead (2009). In the present study, the smiles are segmented into the three phases of the smile and the durations of each phase is used to train the Support Vector Machines.

Characteristics of spontaneous and posed smiles analyzed in the literature are often used for the automatic classification of smiles. Cohn and Schmidt (2004) propose a linear discriminant classifier that classifies posed and spontaneous smiles using information about their amplitude, duration and the relation between the amplitude and duration. Smile amplitude information is also used in the present study.

Valstar et al. (2007) not only focus on facial information, but propose a multimodal approach fusing video data from the face, head and shoulders. The resulting classifier is very accurate and it transpires that head motion is the most reliable source of data, followed by the face. Hence, head motion is used to train the models in the current study and, furthermore, models trained on several sources of data, head motion included, are fused to investigate the multimodal approach for posed and spontaneous enjoyment smile classification.

In contrast to the multimodal approach of Valstar et al. (2007), Dibeklioğlu et al. (2010) propose a classifier that merely makes use of eyelid movements and propose distance-based and angular features for eyelid movements. This classifier is compared to classifiers trained on other facial features and the results indicate that eyelid movements are more reliable for smile classification than movements of the eyebrows, cheeks and lips. In addition, the results show that lip motion is the second most reliable source of information for smile classification. The present study therefore investigates, besides the influence of temporal information from the head, the influence of temporal information from the eyes and the lips on the classification of enjoyment smiles. In addition, the trained classifiers in the study by Dibeklioğlu et al. (2010) include Continuous Hidden Markov Models, which are also used in the present study. However, in the study by Dibeklioğlu et al. (2010) data from different facial regions are used to train seperate models and are not fused for improved classification. This fusion is performed, however, in the present study for data from the eyelids, the lips and the head. Furthermore, the present study uses different lip features than the ones presented by Dibeklioğlu et al. (2010).

More recently, Dibeklioğlu et al. (2012) propose a classifier that distinguishes between spontaneous and posed enjoyment smiles by using the dynamics of eyelid, cheek, and lip corner movements. Support Vector Machines are trained on different features for different phases of the smile and it is shown that for different phases of the smile, different facial regions are more descriptive. Although this classifier makes use of dynamic features, the dynamic temporal information of the face is more implicitly preserved than in a Continuous Hidden Markov Model (CHMM). Whereas Continuous Hidden Markov Models can be trained on raw sequences of, for example, lipcorner amplitude values, Support Vector Machines require less raw training data with fewer dimensions. This is for example achieved by taking the mean of a sequence of lipcorner amplitudes in one of the smile phases. However, such an approach basically ignores most temporal information. Dibeklioğlu et al. (2012) take a different approach and computed the mean, maximum and standard deviation for segments with increasing and decreasing values in each of the three phases of the smile (and for different regions of the face). In addition, speed and acceleration features are computed. Therefore, although temporal information is generally less explicitly preserved in a SVM than in a CHMM, SVMs can be used to investigate the influence of temporal facial information when the features are computed in a way guaranteeing the preservation of a reasonable amount of temporal information. Hence, in the present study both CHMMs and SVMs are trained to distinguish between posed and spontaneous enjoyment smiles.

In summary, based on existing literature the present study investigates the influence of temporal information from the head, lips and eyes, and the influence of the fusion of these sources of information, on the classification of posed and spontaneous enjoyment smiles using CHMM and SVM classifiers.

3 Method

In order to investigate the influence of temporal facial information on the classification of posed and spontaneous enjoyment smiles, data is gathered and preprocessed, features are extracted and models are trained and then fused. A visualisation of this pipeline is shown in Figure 2. The sources of temporal information that are investigated are head motion, lip movement and eye movement.

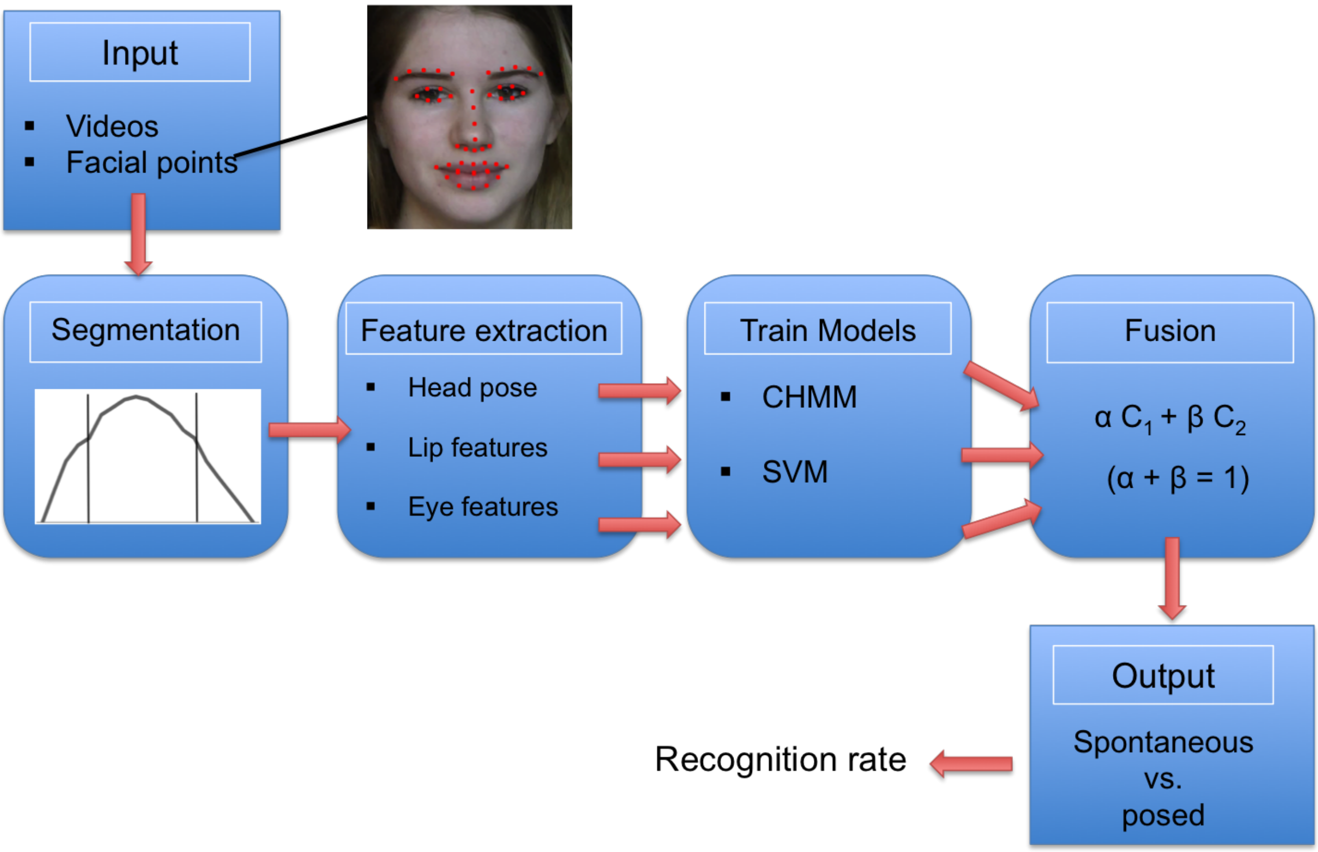

Figure 2: The pipeline of the research method

3.1 Registration

Smile videos of varying lengths are used, see Section 4.1 for more details about the dataset. In each video 49 facial landmarks are initialised in the first frame and then tracked throughout the video. The facial landmarks include points of the eyes, eyebrows, nose and lips, and are shown in Figure 3. The landmarks are further denoted as l_i^t, which means the i'th landmark in frame t. The facial landmarks and head pose angles are estimated using the Supervised Descent Method (SDM) proposed by Xiong and De la Torre (2013). Due to the usage of SIFT features extracted from patches around the landmarks, a robust representation against illumination is obtained. Hence, because the estimation of the landmarks is robust against illumination, the classifier trained on this data is also robust against illumination, satisfying the criterion of illumination robustness presented in Section 1.

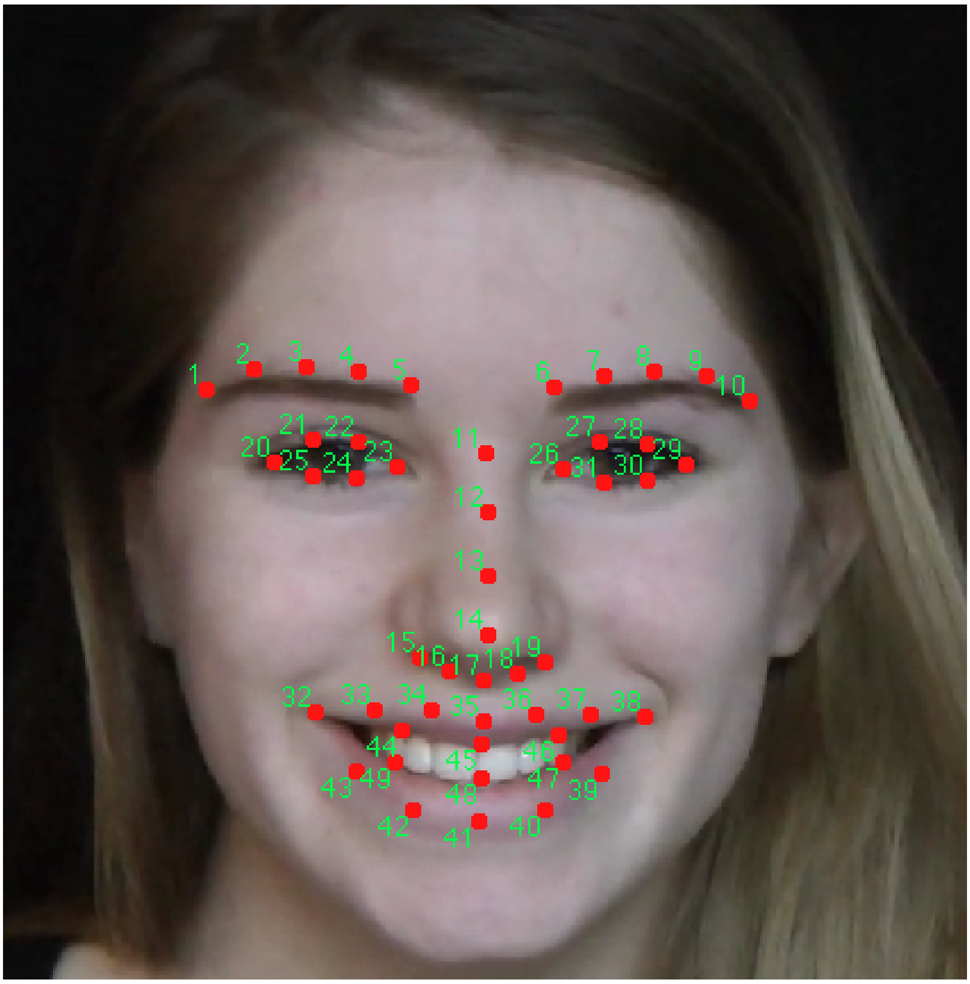

Figure 3: The 49 facial landmarks

3.2 Temporal segmentation

To satisfy the criterion of free smile duration (see Section 1), videos with smiles of varying lengths are used for training. To align the onset, apex and offset phases of the smiles in the videos, the videos are temporally segmented.

Onset, apex and offset phases of the smile can be detected by computing lip corner displacement for every frame and selecting the initial longest continuous increase as onset, the longest continuous decrease as offset and the timespan between the onset and offset as apex (see Figure 4). Here, lip corner displacement is computed by averaging left and right lip corner displacement vectors and polynomials are fit to the resulting function for smoothing purposes. Subsequently, onset, apex and offset are selected on the smoothed curve.

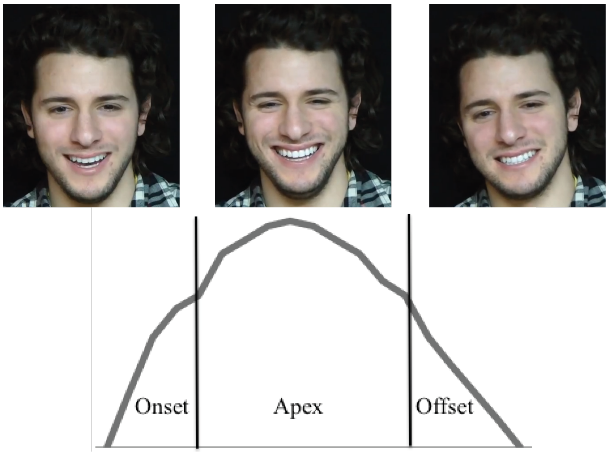

Figure 4: The temporal segmentation of a smile into the three phases onset, apex and offset.

3.3 Feature extraction

The temporal segmentation of the smiles is followed by feature extraction at different regions: the head, the lips and the eyes. These features are extracted from each phase of the smile.

3.3.1 Head pose

There are three angles that determine the head pose: roll, yaw and pitch (see Figure 5). The angles are defined within a range of $-180°$ to $+180°$, where a value of $0°$ for each of the angles corresponds to a neutral head pose. As previously mentioned in Section 3.1, the roll, yaw and pitch angles are the output of the landmarker by Xiong and De la Torre (2013). Each head pose signal is smoothed with (4253H, twice) (Velleman, 1980), which is a running median smoother of 42, 5 and 3. Subsequently, the signal is normalized with respect to the first frame by subtracting the first frame from the signal.

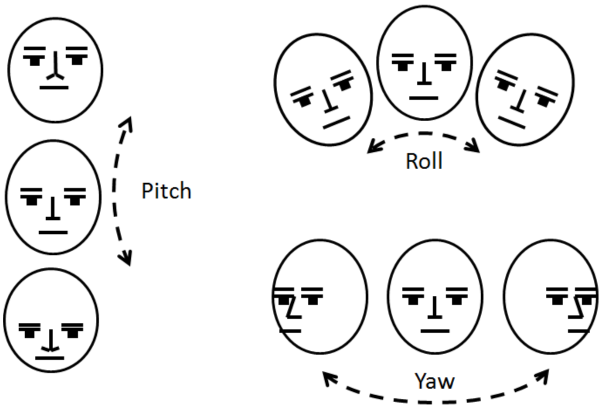

Figure 5: Head pose angles. Retrieved from `http://msdn.microsoft.com/en-us/library/jj130970.aspx`

3.3.2 Lip features

The principal indicator of a smile, either posed or spontaneous, is the raising of the lip corners, caused by the contraction of the zygomaticus major muscle. The distance between the center of the mouth and the lip corners (smile amplitude) increases during the onset of a smile, until the maximum amplitude is reached in the apex phase, after which the amplitude decreases during the offset phase. In addition, the angle of raised lip corners with respect to their non-raised pose also changes during a smile.

Hence, two different lip features are used: two-dimensional lip corner displacement features and lip corner angle features. The facial landmarks used for the computation of these lip features are normalized with respect to rotation, translation and scale. This satisfies the criteria of invariance to scale and head movements presented in Section 1. Furthermore, once the lip features are computed, the signals are smoothed with the running median smoother (4253H, twice) (Velleman, 1980). In addition, the lip corner angle signals are normalized with respect to the first frame by subtracting the feature value of the first frame from the entire signal.

3.3.2.1 Lip corner displacement vector

The lip corner displacement feature $\mathcal{D}_{lipdisp}$ is a vector from the center of the lips to the corners of the lips (see Figure 6). Because $\mathcal{D}_{lipdisp}$ is a vector it encodes both magnitude (length) and direction. This is useful because, in addition to the changing lip corner angle during a smile, the distance between the center of the lips and the lip corners also changes. This change in distance is further referred to as lip corner displacement.

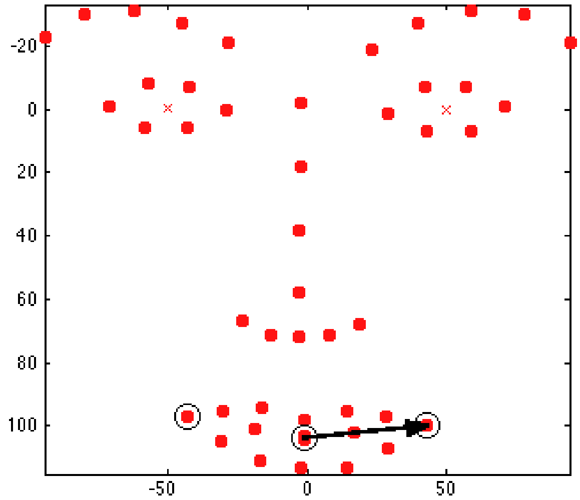

Figure 6: Left lip corner displacement vector on a normalized face.

The lip corner displacement feature $\mathcal{D}_{lipdisp}$ is computed for the right and left lip corner by subtracting the center of the mouth from the lip corners and normalizing by the length of the lip (Euclidian distance between the lip corners) in the first frame. The length of the lip in the first frame is used for the normalization of the extracted features in the rest of the frames because the Euclidian distance between the lip corners changes during a smile and will, therefore, not correspond to the length of the lip in most frames. The computation of the right lip corner displacement feature $\mathcal{D}_{lipdisp_R}(t)$ and left lip corner displacement feature $\mathcal{D}_{lipdisp_L}(t)$ are as follows:

$$\mathcal{D}_{lipdisp_R}(t) = \frac{(l_{32}^t - \frac{l_{45}^t + l_{48}^t}{2})}{d(l_{32}^1, l_{38}^1)}, \tag{1}$$

$$\mathcal{D}_{lipdisp_L}(t) = \frac{(l_{38}^t - \frac{l_{45}^t + l_{48}^t}{2})}{d(l_{32}^1, l_{38}^1)}. \tag{2}$$

As shown in Figure 7, facial landmarks l_{32}^t and l_{38}^t correspond to the right and left lip corner in frame t, respectively. In addition, $\frac{l_{45}^t + l_{48}^t}{2}$ corresponds to the center of the mouth in frame t.

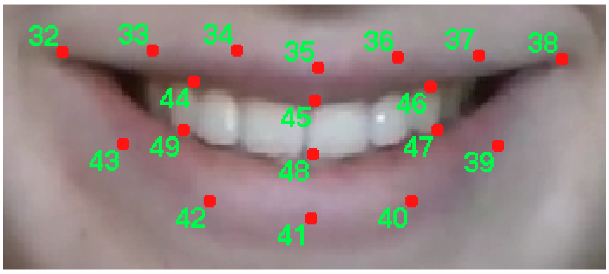

Figure 7: The landmarks on the lips

3.3.2.2 Lip corner angle

The lip corner angle feature is defined as the angle between the raised lip corners and their non-raised counterparts. However, due to normalization of the facial landmarks prior to the computation of the features, the lip corner angles can be computed by calculating the angle between the raised right and left lip corner vectors and the unit vectors $\left[\begin{smallmatrix}-1\\0\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, respectively (see Figure 8, where the unit vectors are stretched for clarity reasons). The raised right and left lip corner vectors are identical to the lip corner displacement vectors in the previous section, see Equation 1 and 2 their computation, and therefore the lip corner vectors will be further denoted as $\mathcal{D}_{lipdisp_R}(t)$ and $\mathcal{D}_{lipdisp_L}(t)$.

In general, the angle α between two vectors $\mathbf{A}$ and $\mathbf{B}$ can be computed as follows:

$$\alpha = \cos^{-1}\left(\frac{\mathbf{A} \bullet \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}\right).$$

Hence, the angles $\mathcal{D}_{lipangle}$ (in degrees) for the right and left lip corner can be computed as follows:

$$\mathcal{D}_{lipangle_R}(t) = \cos^{-1}\left(\frac{\left[\begin{smallmatrix}-1\\0\end{smallmatrix}\right] \bullet \mathcal{D}_{lipdisp_R}(t)}{\left\|\left[\begin{smallmatrix}-1\\0\end{smallmatrix}\right]\right\| \cdot \|\mathcal{D}_{lipdisp_R}(t)\|}\right), \tag{3}$$

$$\mathcal{D}_{lipangle_L}(t) = \cos^{-1}\left(\frac{\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right] \bullet \mathcal{D}_{lipdisp_L}(t)}{\left\|\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\right\| \cdot \|\mathcal{D}_{lipdisp_L}(t)\|}\right). \tag{4}$$

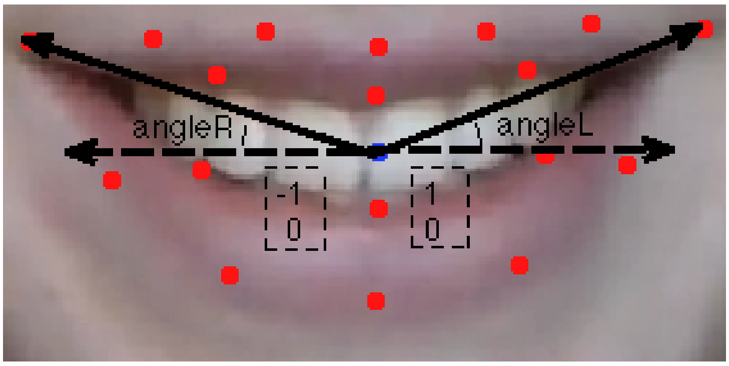

Figure 8: The lip corner angle features (right and left).

3.3.3 Eye features

Two different eye features are used: eyelid displacement and an eye aperture angle. Before the extraction of eye features, all faces are normalized with respect to rotation, translation and scale. This satisfies the criteria of invariance to scale and head movements. Furthermore, once the eye features are computed the signals are smoothed with the running median smoother (4253H, twice) (Velleman, 1980). In addition, the signals are normalized with respect to the first frame by subtracting the feature value of the first frame from the feature values of all frames.

3.3.3.1 Eyelid displacement

The eyelid displacement feature $\mathcal{D}_{eyelid}(t)$ for frame t is defined as the Euclidian distance between the middle point of the eye $P_{0,3}$ and the middle point of the upper eye curve $P_{1,2}$ in frame t, normalized by the length of the eye (see Figure 9). Firstly, the middle point of each eye $P_{0,3}$ is computed by taking the mean of the corner points P_0 and P_3 of the eye.

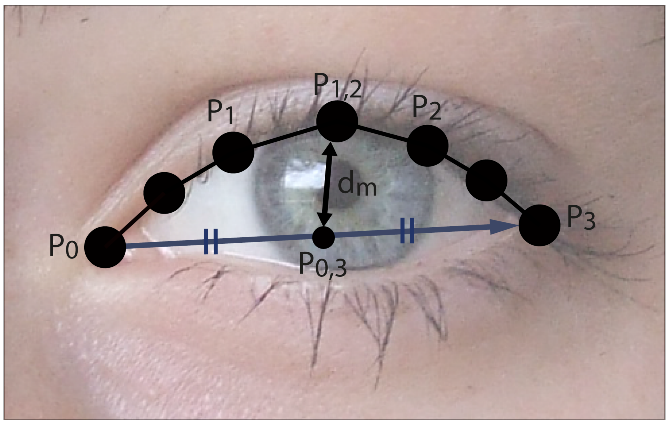

Figure 9: The eyelid displacement feature $\mathcal{D}_{eyelid}(t)$ (d_m in figure).

As can be seen in Figure 3, the middle point of the upper eye curve is not one of the 49 facial landmarks. Therefore, in order to find the middle point of the upper eye curve $P_{1,2}$, a cubic Bézier curve is fitted to the facial landmarks l_{20}^t to l_{23}^t for the right eye and to facial landmarks l_{26}^t to l_{29}^t for the left eye. As can be seen in Figure 3 these landmarks are the four upper landmarks for each eye. The cubic Bézier curve through four points $P_0 ... P_3$ is defined by:

$$B(b) = (1-b)^3 P_0 + 3(1-b)^2 b P_1 + 3(1-b) b^2 P_2 + b^3 P_3, \quad b \in [0,1]. \tag{5}$$

Here, B(b) corresponds to P_1 and P_2 for $b = \frac{1}{3}$ and $b = \frac{2}{3}$, respectively. Hence, B(b) corresponds to $P_{1,2}$ for $b = \frac{1}{2}$. Hence, the middle point of the upper eye curve for the left and right eye for frame t can be computed as follows:

$$B_R^t\left(\frac{1}{2}\right) = \frac{1}{8} l_{20}^t + \frac{3}{8} l_{21}^t + \frac{3}{8} l_{22}^t + \frac{1}{8} l_{23}^t,$$

$$B_L^t\left(\frac{1}{2}\right) = \frac{1}{8} l_{26}^t + \frac{3}{8} l_{27}^t + \frac{3}{8} l_{28}^t + \frac{1}{8} l_{29}^t.$$

Furthermore, the length of the eye is defined as the Euclidian distance between the corner points of the eye $d(P_0, P_3)$, where $d()$ denotes the Euclidian distance. This eye length is used for normalization.

Finally, the eyelid displacement feature for the right eye $\mathcal{D}_{eyelid_R}(t)$ and left eye $\mathcal{D}_{eyelid_L}(t)$ are thus defined as follows:

$$\mathcal{D}_{eyelid_R}(t) = \frac{d\left(\frac{l^t_{20}+l^t_{23}}{2}, B^t_R(\frac{1}{2})\right)}{d(l^t_{20}, l^t_{23})},$$

$$\mathcal{D}_{eyelid_L}(t) = \frac{d\left(\frac{l^t_{26}+l^t_{29}}{2}, B^t_L(\frac{1}{2})\right)}{d(l^t_{26}, l^t_{29})}.$$

3.3.3.2 Eye aperture angle

There are two eye aperture angle features, one computed using a landmark on the eyelid estimated using the Bézier curve fitted to the recorded landmarks on the eyelid, $\mathcal{D}_{eyeangleB}(t)$, and the other computed using exclusively the recorded landmarks, $\mathcal{D}_{eyeangle}(t)$. See Figure 10 and Figure 11 for the two eye aperture angle features.

The Bézier eye aperture angle feature $\mathcal{D}_{eyeangleB}(t)$ is defined as the angle between the vector $\mathbf{v}_2$, stretching from one corner to the other corner of the eye, and vector $\mathbf{v}_1$, stretching from the outer corner of the eye P_0 to the neighbouring landmark $P_{0,1}$ (see Figure 10). In Figure 3, it is shown that landmark $P_{0,1}$ is not a recorded landmark and, thus, needs to be estimated. This is done by fitting a Bézier curve $B(b)$ to the recorded landmarks P_0 to P_3 so that $P_{0,1} = B^t_R(\frac{5}{6})$ for the right eye and $P_{0,1} = B^t_L(\frac{1}{6})$ for the left eye. (see Equation 5 for $B(b)$). The Bézier eye aperture angle features $\mathcal{D}_{eyeangleB_R}(t)$ and $\mathcal{D}_{eyeangleB_L}(t)$ for the right and left eye, respectively, can then be computed as follows:

$$\mathcal{D}_{eyeangleB_R}(t) = \cos^{-1}\left(\frac{(B^t_R(\frac{5}{6}) - l^t_{23})\bullet(l^t_{23} - l^t_{20})}{\left\|B^t_R(\frac{5}{6}) - l^t_{23}\right\| \cdot \|l^t_{23} - l^t_{20}\|}\right), \tag{6}$$

$$\mathcal{D}_{eyeangleB_L}(t) = \cos^{-1}\left(\frac{(B^t_L(\frac{1}{6}) - l^t_{26})\bullet(l^t_{26} - l^t_{29})}{\left\|B^t_L(\frac{1}{6}) - l^t_{26}\right\| \cdot \|l^t_{26} - l^t_{29}\|}\right). \tag{7}$$

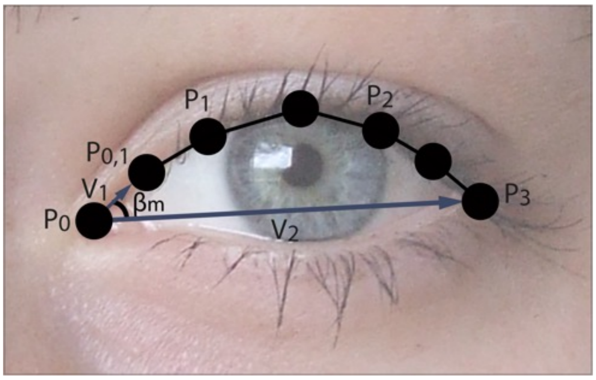

Figure 10: The Bézier eye aperture angle feature $\mathcal{D}_{eyeangleB}(t)$ (β_m in figure).

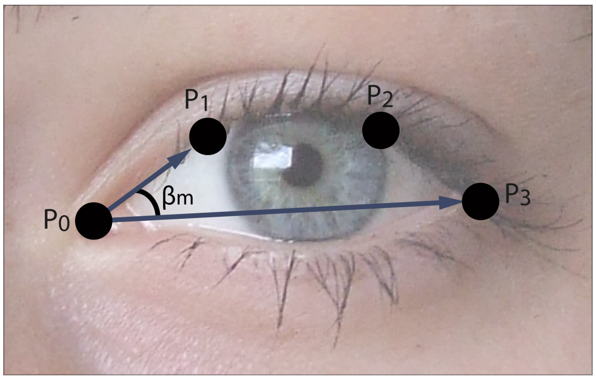

Figure 11: The eye aperture angle feature $\mathcal{D}_{eyeangle}(t)$ (β_m in figure).

The eye aperture angle feature $\mathcal{D}_{eyeangle}(t)$, which is defined exclusively by recorded landmarks, can be computed quite similarly to $\mathcal{D}_{eyeangleB}(t)$, namely, by replacing $B^t_R(\frac{5}{6})$ and $B^t_L(\frac{1}{6})$ in Equations 6 and 7 by l^t_{22} and l^t_{27}, respectively.

3.4 Models

Modelling temporal facial information requires a model that can handle temporal information. Because temporal facial information can be represented as a sequence of symbols it is natural to consider a Hidden Markov Model (HMM). The advantage of an HMM is that it accepts sequences with variable lengths, which is useful because the videos are of varying lengths. Moreover, this satisfies the criterion of free smile duration.

Rather than the more commonly known and used discrete Hidden Markov Model, the continuous variant is used: a continuous Hidden Markov Model (CHMM). In contrast to a DHMM, a CHMM accepts sequences of any value on the real line. This is useful for modelling temporal facial information because temporal information can be represented as a sequence of real values, where each consecutive value represents the information of the next frame. The (continuous) Hidden Markov Model is further explained in Section 3.4.1.

In addition to continuous Hidden Markov Models, Support Vector Machines (SVM) are used. As opposed to the probabilistic CHMM, a SVM is a non-probabilistic classifier. A Support Vector Machine is a binary classifier in which each sample is labeled with its class and represented as an n-dimensional feature vector in high-dimensional space. During training, the SVM attempts to map the feature vectors to points in high-dimensional space in such a way that the best separation between the two classes is achieved, which is, a clear gap between the two classes. Although SVMs require feature vectors (samples) to be of equal length, SVMs can be used for the classification of smiles with varying durations by uniformly segmenting them and extracting features from the resulting segments (see Section 3.4.2 for more details).

3.4.1 Hidden Markov Model

A Hidden Markov Model is a Markov chain in which the states are unobserved or hidden. A Markov chain is a system consisting of states in which the transition from one state to another exclusively depends on the current state and not on the preceding state transitions. A Markov chain can be used to compute a probability for a sequence of events observable in the world. However, in smile spontaneity classification the event of interest, whether a person is smiling spontaneously or is posing a smile, is not directly observable. The sequences of features extracted from the smile, however, are observable.

A Markov chain cannot be used to model a hidden process based on observable events, but a Hidden Markov Model can. Hence, a Hidden Markov Model can be used for smile spontaneity classification, where the intention of the smile (spontaneous or posed) is hidden, but the facial actions during the smile are observable. In Table 1 the correspondence between smiles and the elements of the HMM can be seen.

	Smile	**HMM**
Hidden	Posed/spontaneous	State in model
Observable	Facial action	Sequence of symbols
Temporal domain	Dynamic behaviour	A network of state transitions
Characteristics	Smile	State transition probability and symbol probability
Recognition	Smile similarity	The confidence of output probability

Table 1: The correspondence between smiles and elements of the HMM. Adopted from Lien (1998).

More formally, an HMM can be defined as a triple $\lambda = (\Pi, A, B)$. Here, $\Pi = (\pi_i)$ is the vector of initial state probabilities, which are the probabilities of a sequence starting in a particular state. Furthermore, $A = (a_{ij})$ is the state-transition matrix, where (a_{ij}) is the probability of moving from state i to state j. Finally, $B = (b_{ij})$ is the observation probability matrix, containing the probabilities of observing an event o_i in state x_j.

The classification of posed and spontaneous enjoyment smiles requires two HMMs, one for each class. During training, characteristics are extracted from each sample in the train set, mapped to the corresponding observation sequence O and the model parameters (Π, A, B) are adjusted in a way that maximizes $P(O|\lambda)$, the probability of the observed sequence given the model. For the classification of an unseen sample, $P^w = P(O|\lambda^w)$ is computed for both HMMs. Subsequently, the class w corresponding to the model with the maximum likelihood is selected as the final classification result.

3.4.1.1 Continuous Hidden Markov Model

The principal difference between a discrete HMM and a continuous HMM is that a continuous HMM accepts continuously varying feature sequences and a discrete HMM constrains the feature sequences to take values from a discrete finite alphabet. For this reason a CHMM rather than a DHMM is used for modeling the continuously varying smile feature sequences.

Due to its continuous nature, a CHMM does not have a discrete set of prior state probabilities π_i, state-transition probabilities a_{ij} and observation probabilities b_{ij}, like a DHMM[2], but a CHMM has mixtures of multivariate of Gaussians as probability distribution functions. During training, the mean and variance of each of the Gaussians is optimized to maximize the probability of the observed sequence.

3.4.2 Support Vector Machine

As previously mentioned in Section 3.4, a Support Vector Machine is a binary classifier, which means that it classifies samples into two classes. Each sample is an n-dimensional feature vector which is mapped to a data point in high-dimensional space and, subsequently, during training an $(n-1)$-dimensional hyperplane is found that best separates the data points of the two classes.

When the data points are linearly separable, they can be linearly separated by two hyperplanes in a way that the hyperplanes have no data points in between them. Then, a third hyperplane, the maximum-margin hyperplane, can be selected in between the two hyperplanes that maximizes the distance to the two hyperplanes (see Figure 12).

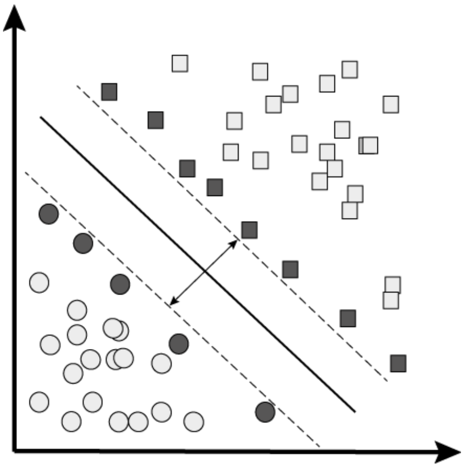

Figure 12: The maximum-margin hyperplane separating two classes.

[2] Although DHMMs always have a discrete set of prior state probabilities and state-transition probabilities, the observation probability distribution can either be discrete or continuous.

However, when the data points are not linearly separable by a hyperplane, other kernel functions can be used to map the feature vectors to data points in high-dimensional space in a different way in order that they are linearly separable in high-dimensional space. The radial basis function (RBF) is an example of such a kernel function (see Figure 13). RBF nonlinearly maps the feature vectors of the samples into high-dimensional feature space, where the points can be linearly separated by a hyperplane (see Figure 14). The linear kernel and the RBF kernel will be used in the present study.

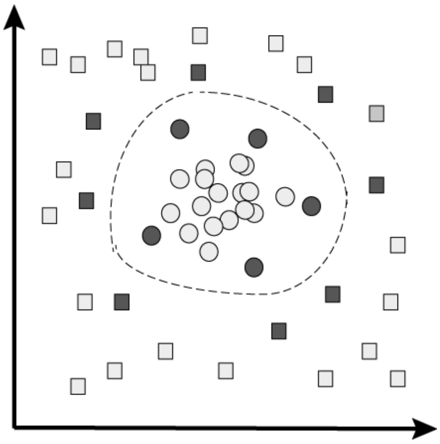

Figure 13: The radial basis function.

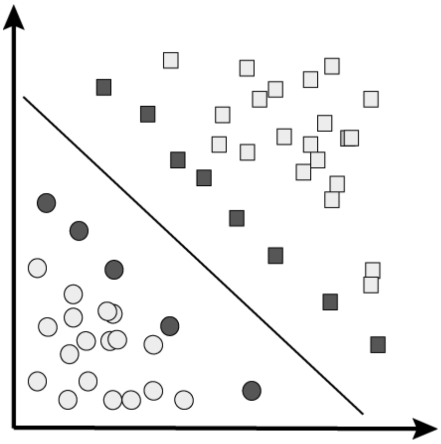

Figure 14: RBF mapping.

3.4.2.1 Feature vector computation

In order to use the extracted features described in Section 3.3 for the training of SVMs, they need to be altered because the sequences of feature values are of varying lengths due to the varying lengths of the smiles. In addition, a sequence of values for one feature for one phase can be 100 frames long, which would result in a 100-dimensional vector for a single feature for a single phase.

These problems can be avoided by selecting a single value, for example the mean, to represent an entire sequence. This way, the number of dimensions is greatly reduced to three dimensions per feature (the mean value of the onset, apex and offset). However, this approach ignores a great amount of temporal information within each phase. Therefore, a different approach is taken in the present study.

Firstly, the mean signal is computed from the left and right feature signals (for features that have a left and right variant). Then, for every phase of each feature, increasing and decreasing segments are detected in the feature signal. For the onset phase of the eye aperture angle feature, for example, this means that increasing and decreasing segments are detected in which the eye aperture angle is increasing and decreasing, respectively (see Figure 15). Subsequently, the mean, standard deviation, maximum value and duration ratio of the increasing segment group and decreasing segment group are computed, which results in eight values per feature per phase. The duration ratio of a segment group is computed by dividing the number of frames in the segment group by the total number of frames in the increasing and decreasing segment group.

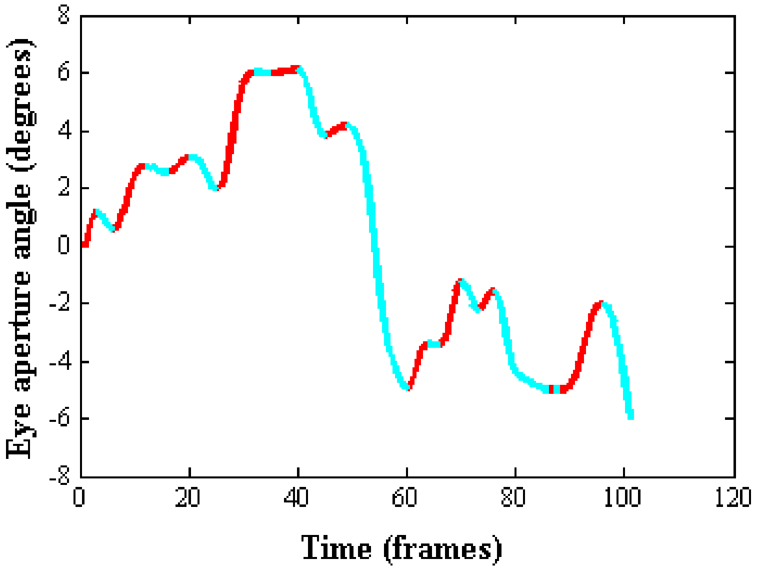

Figure 15: Selection of increasing (red) and decreasing segments (blue) of eye aperture angles (normalized with respect to the first frame) for the onset of a sample in the dataset.

Furthermore, the first and second derivatives of the feature signals were computed to obtain speed and acceleration signals for each feature[3]. The increasing and decreasing segments are detected and the mean and maximum of each segment group is computed. Subsequently, for each feature the mean, standard deviation, maximum and duration ratio of the amplitude (regular) signals are concatenated with the mean and maximum of the speed and acceleration signals.

In order to add more temporal information, the duration of each phase of the smile is computed. This feature is not added to the separate feature vectors because it interferes with the results for the separate features, but it is added to the concatenation of all the feature vectors to show its effect.

Lastly, the feature vectors are normalized with min-max normalization, so that they are rescaled between 0 and 1. The principal advantage of normalization is that it standardizes the range of the features so that features with values in greater ranges do not dominate the features with values in smaller ranges.

3.5 Fusion

Different features may be more discriminating in different phases of the smile. In addition, certain combinations of features may be more discriminating than the features separately. Therefore, fusion of models trained on different features and phases may improve classification.

CHMMs trained on different features can be fused either by fusing the maximum likelihood estimates or by majority voting. In the first approach, the maximum likelihood of the observed sequence P_i^w is computed for each model λ_i. Subsequently, the fused maximum likelihood P_f of N models are computed for each class w by computing a linear combination of the maximum likelihoods P^i as follows:

$$P_f^w = \sum_{i=1}^{N} a_i P_i^w, \quad \sum_{i=1}^{N} a_i = 1.$$

It should be noted that in the present study the weights a_i are equal to each other. The weights could, however, be optimized by learning.

[3] Speed and acceleration signals are not computed for the lip displacement vector features.

The second approach to fuse CHMMs is majority voting of the final classification results. This means that for each CHMM pair i (the model trained on the posed smiles and the model trained on the spontaneous smiles) the class w is voted for where $P^w = maxP_i^w(O|\lambda_i)$ and, subsequently, the votes are accumulated for each class and the class with the majority of the votes is selected as the final classification result.

Majority voting is not used in this study, because it is more likely that the fusion of maximum likelihoods gives more accurate results. This can be illustrated in the following example. Three models trained on different features are fused. These three models will output two maximum likelihood estimates for each class (per sample). For sample i, the first model outputs a maximum likelihood estimate of 49.9% for class 1 and 50.1% for class 2. The second model outputs a maximum likelihood estimate of 49% for class 1 and 51% for class 2. Apparently, the first and second model cannot classify the sample well, as the maximum likelihood estimates are very close to each other. The third model outputs a maximum likelihood estimate of 96% for class 1 and 4% for class 2 for this sample. Hence, the model is quite confident the sample should be classified as class 1[4]. Then, maximum likelihood fusion results in 64.97% for class 1 and 35.03% for class 2. Hence, with maximum likelihood fusion, sample i is classified as class 1.

However, with majority voting, the first and second model vote for class 2 and the third model votes for class 1, which results in the sample being classified as class 2. As can be noticed, the two models that classify the sample with low confidence overshadow the model that classifies the sample with high confidence. Therefore, maximum likelihood fusion is preferred over majority voting and, hence, maximum likelihood fusion is used in this study to fuse CHMMs. This type of fusion is called late fusion.

As opposed to the late fusion scheme used for the CHMM classifiers, an early fusion scheme is used for the SVM classifiers. This type of fusion does not include the fusion of model parameters or majority voting, but the fusion of the data used to train and test the model. For example, the fusion of the SVM trained on head pose features and the SVM trained on eye features can be fused by training a third SVM on the concatenation of the head pose and eye features.

[4] The high confidence that the sample belongs to class 1 does not necessarily imply that the sample does belong to class 1. However, it is likely that this is the case.

4 Experiments

For the classification of spontaneous and posed smiles, individual CHMM and SVM classifiers are trained on different features from the head, lips and eyes. For the CHMM classifiers 10-fold cross-validation is used to train and test the models trained with different numbers of states. The model that performs best is used for fusion[5]. The number of gaussians in the mixtures is set at six for most models, but for CHMM classifiers trained on lip features eight gaussians in the mixtures is found to perform better.

Furthermore, for the SVM classifiers, a two-level 10-fold cross-validation scheme is used, where each time a test fold is left out, 5-fold cross-validation is used to train the system and optimize the parameters. SVM classifiers with RBF (with parameter selection) are found to have a better performance than classifiers with a linear kernel when trained on individual features (see Appendix A.2). However, classifiers with a linear kernel perform slightly better when trained on all features (see Appendix D.1) . Hence, the focus of Section 4.3 is on the results of SVM classifiers with RBF rather than on the results for linear SVM classifiers.

4.1 Dataset

The dataset employed for the experiments is the UvA-NEMO Smile Database (Dibeklioğlu et al., 2012). This database contains 1240 smile videos (597 spontaneous, 643 posed) from 400 subjects (185 female, 215 male) with ages varying from 8 to 76 years. Age and gender distributions for the subjects and smiles can be seen in Figure 16. To obtain the posed enjoyment smiles subjects were instructed to pose a smile as realistically as possible. Furthermore, for the spontaneous enjoyment smiles short and funny videos were used to elicit smiles.

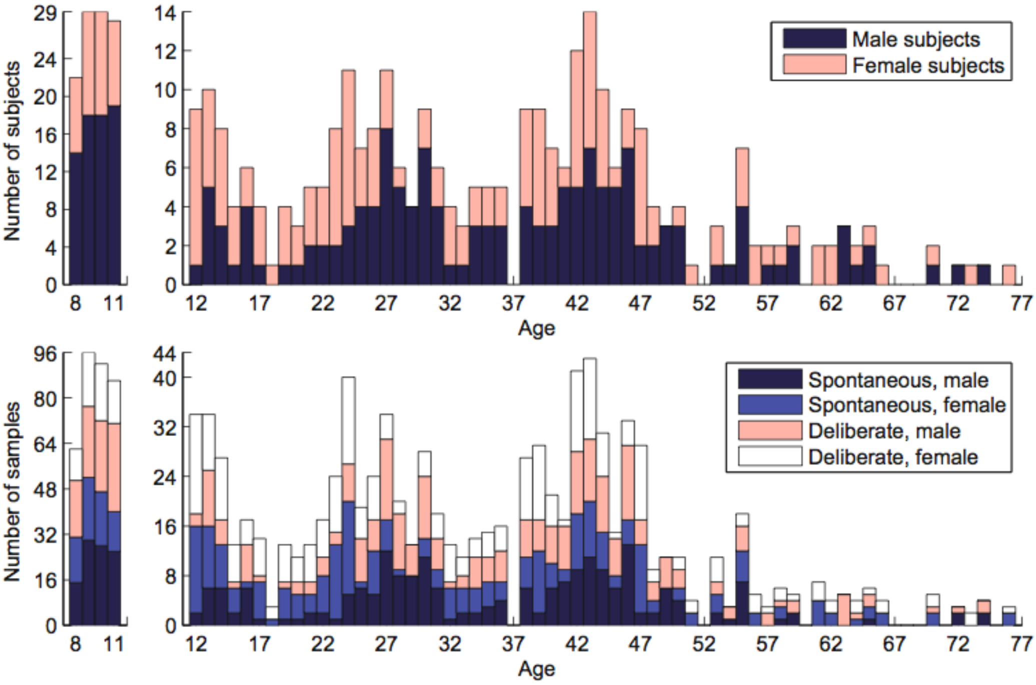

Figure 16: Age and gender distributions for the subjects (top), and for the smiles (bottom) in the UvA-NEMO Smile Database. Adapted from Dibeklioğlu et al. (2012).

[5] For each feature CHMM classifiers with a number of states varying from one state to ten states are trained. The CHMM with the highest mean recognition rate (of the onset, apex, offset and the entire smile) is selected for fusion with CHMM classifiers trained on other features.

The videos are recorded in RGB colour with a resolution of 1920×1080 pixels with a frame rate of 50 frames per second. In addition, the videos are recorded under controlled illumination conditions and a colour chart is present on the background of the videos (see Figure 17). Furthermore, each video starts and ends with a neutral or near-neutral expression. For more details about the database see (Dibeklioğlu et al., 2012).

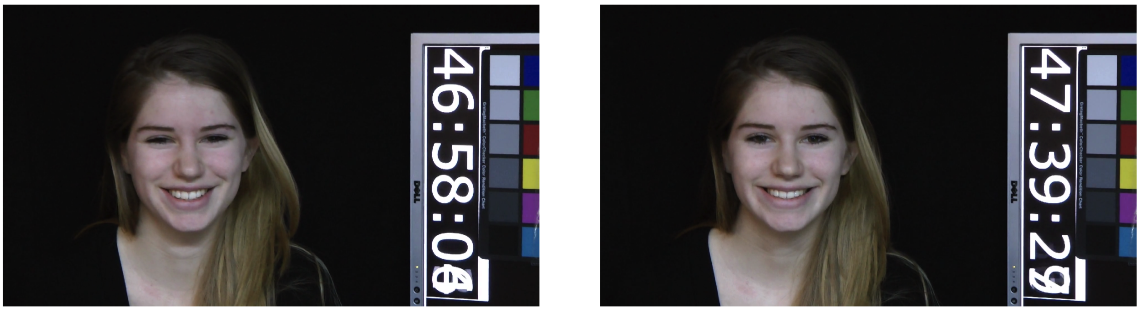

Figure 17: Spontaneous (left) and posed (right) enjoyment smiles from the UvA-NEMO Smile Database

4.2 Metrics

The metric that is used to evaluate the models is the recognition rate. For both the CHMM and SVM classifiers, the recognition rate is the combined accuracy of the classification for both classes. For example, in the confusion matrix in Table 2, the recognition rate is computed as follows (here, $a + b = c + d$, which is the size of the test set):

$$Rec = \frac{\frac{a}{a+b} + \frac{d}{c+d}}{2} \cdot 100\%$$

		Classified as:	
		Spontaneous	Posed
True class:	Spontaneous	a	b
	Posed	c	d

Table 2: Confusion matrix of spontaneous and posed smile classification.

4.3 Results

CHMM and SVM classifiers are trained on head features, lip features and eye features for different phases of the smile. The results for head pose, lips and eyes are reported in Section 4.3.1, 4.3.2 and 4.3.3, respectively. Furthermore, the results for the fusion of models trained on different regions are reported in Section 4.3.4. Finally, the results of state-of-the-art methods evaluated on the UvA-NEMO Smile Database are reported in Section 4.3.5.

4.3.1 The influence of head pose on classification

To classify posed and spontaneous smiles based on head pose, CHMM and SVM classifiers are trained on the three head pose angles: roll, yaw and pitch. The CHMM classifiers trained on roll, yaw and pitch provide a combined classification and, hence, the CHMM classification results are not divided into separate results for roll, yaw and pitch. In contrast, the SVM classification results comprise the combined as well as the separate results for roll, yaw and pitch.

The results for the head pose CHMM classifiers with the highest recognition rates can be seen in Figure 18. These classifiers have five hidden states. The results for the head pose CHMM classifiers with different numbers of states can be seen in Appendix A.1. As shown in Figure 18, the recognition rates of all phases are just above the level of chance. The three phases combined gives the best result with a recognition rate of 51.68%.

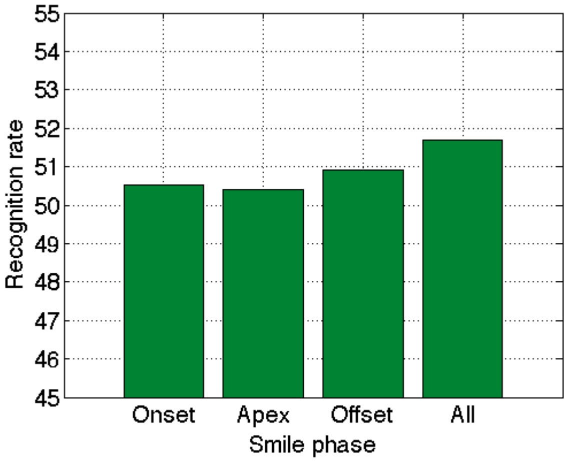

Figure 18: CHMM classification results for head pose for different phases of the smile and the entire smile.

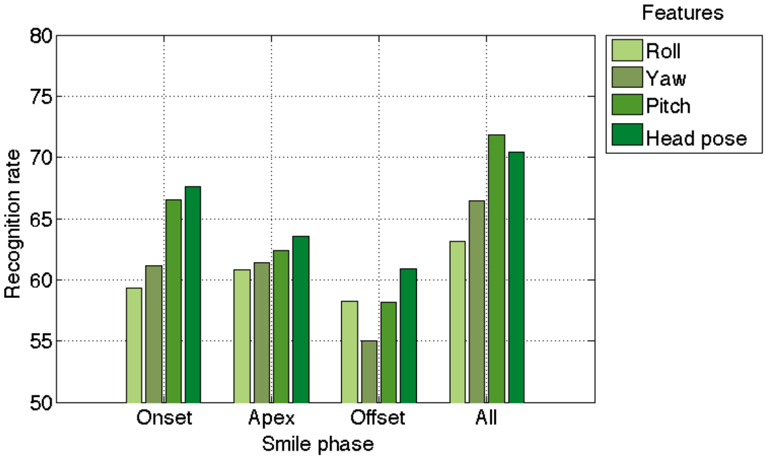

Figure 19: SVM classification results for head pose for different phases of the smile and the entire smile.

The SVM classification results greatly surpass the CHMM results (see Figure 19). In contrast to the low CHMM head pose recognition rate of 51.68% for all phases combined, the corresponding SVM recognition rate of 70.44% is much higher. Pitch angles of all phases combined provides the overall highest recognition rate of 71.88%. Furthermore, each head feature reaches its highest accuracy when the three phases of the smile are combined.

4.3.2 The influence of lip features on classification

To investigate the influence of lip features on the classification of posed and spontaneous enjoyment smiles, CHMM and SVM classifiers are trained on the two lip features: the lip displacement vector features and the lip corner angle features.

The results for the lip CHMM classifiers with the highest recognition rates can be seen in Figure 20. The CHMM classifiers trained on lip displacement features have eight states and the classifiers trained on lip angle features have one state because these are found to have the best performance. The results for lip displacement and lip angle CHMM classifiers with different numbers of states are shown in Appendix B.1 and B.2, respectively. As shown in Figure 20, the combination of both lip features during the apex gives the highest recognition rate for the CHMM classifiers (53.83%). Although this is an improvement of roughly 2% with respect to the head pose CHMM classifiers, it is not a significant improvement because it is still too close to the level of chance.

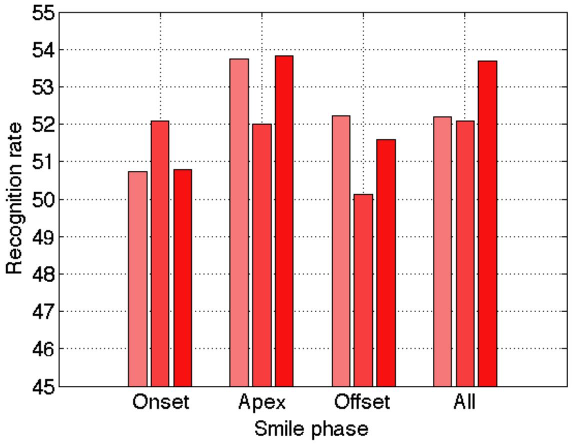

Figure 20: CHMM classification results for lip features for different phases of the smile and the entire smile.

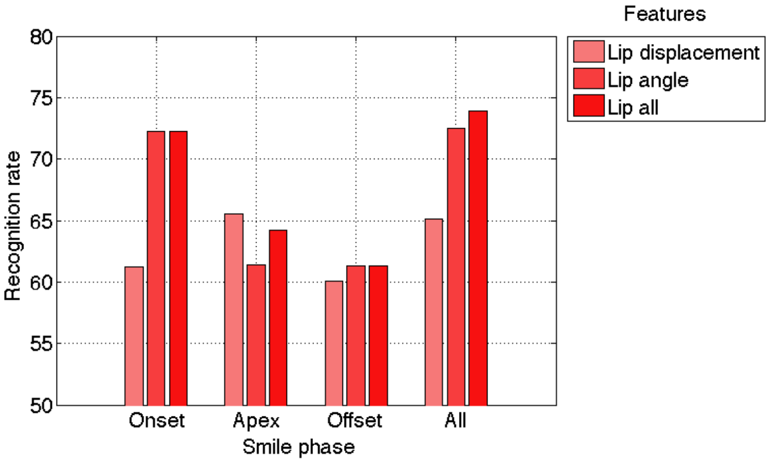

Figure 21: SVM classification results for lip features for different phases of the smile and the entire smile.

SVM classifiers trained on lip features outperform the CHMM classifiers (see Figure 23). Although the highest accuracy for the CHMM classifiers is achieved by combing the lip features during the apex (53.83%), the highest accuracy for the SVM classifiers is achieved by combing the lip features of all phases (73.90%). In comparison to the head pose SVM classifiers (71.88%), this is an improvement of roughly 2%.

Furthermore, the highest accuracy for individual phases is achieved by the onset lip angle features (72.23%). The lip angle features generally achieve a higher accuracy than the lip displacement features, except during the apex where lip displacement (65.57%) exceeds the lip angle features (61.36%) with roughly 4%.

4.3.3 The influence of eye features on classification

To investigate the influence of eye features on the classification of posed and spontaneous enjoyment smiles, CHMM and SVM classifiers are trained on combinations of the three eye features: the Bézier eye aperture angle features, the (non-Bézier) eye aperture angle features and the eyelid displacement features. Subsequently, the combination of eye features that is found to be most discriminating is used for fusion.

The classification results for CHMM classifiers trained on individual eye features and on all eye features combined are shown in Figure 22. The classifiers trained on Bézier eye aperture angle features and (non-Bézier) eye aperture angle features both have six states and the classifiers trained on eyelid displacement features have ten states. The results for Bézier eye aperture angle features, (non-Bézier) eye aperture angle features and eyelid displacement features trained on different numbers of states are shown in Appendix C.1, C.2 and C.3, respectively.

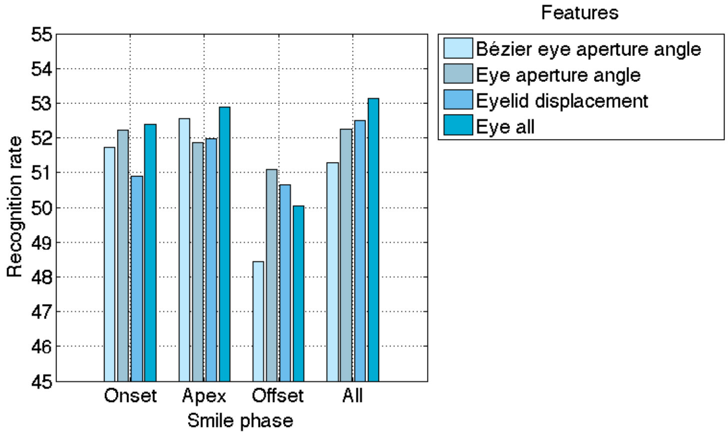

Figure 22: CHMM classification results for eye features for different phases of the smile and the entire smile.

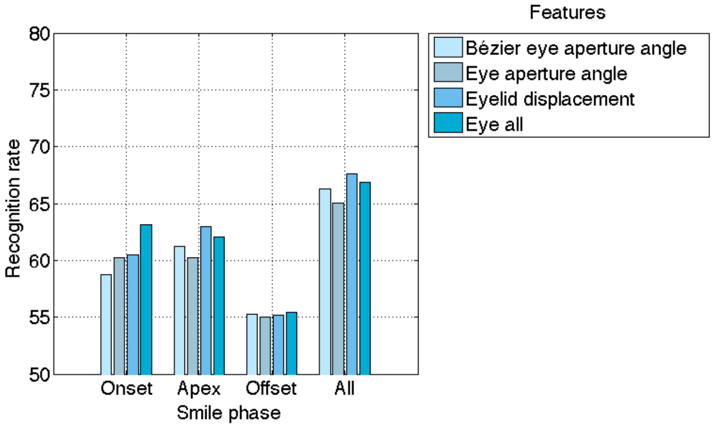

Figure 23: SVM classification results for eye features for different phases of the smile and the entire smile.

The classification results for CHMM classifiers trained on the combination of Bézier eye aperture angle features and eyelid displacement features are shown in Figure 24. The highest accuracy for this combination of features is reached during the apex (52.94%). In Figure 25, the classification results are shown for CHMM classifiers trained on the combination of (non-Bézier) eye aperture angle features and eyelid displacement features. This combination of features is most discriminating during the onset (52.57%). However, the combination of all eye features exceeds both combinations of eye features (when combining all phases) with an accuracy of 53.13% (see Figure 22. Hence, the CHMM classifiers trained on all eye features are selected for the fusion with classifiers trained on head and lip features.

When compared to CHMM classifiers trained on head features (51.68%) and lip features (53.83%), the CHMM classifiers trained on eye features (53.13%) perform better than the head CHMM classifiers, but worse than the lip CHMM classifiers.

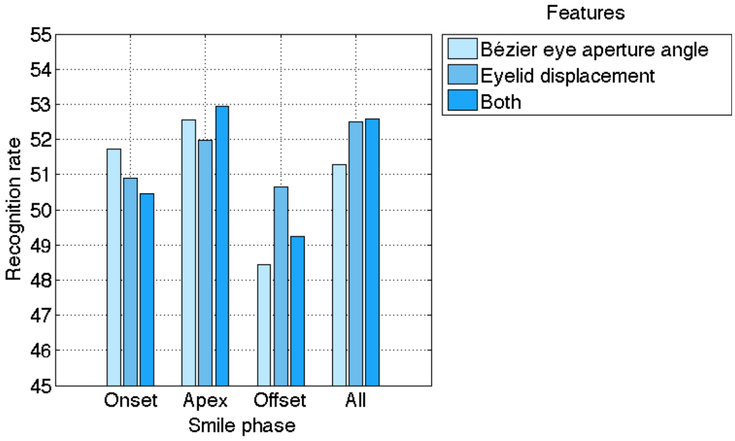

Figure 24: CHMM classification results for Bézier eye aperture angle, eyelid displacement and the two combined.

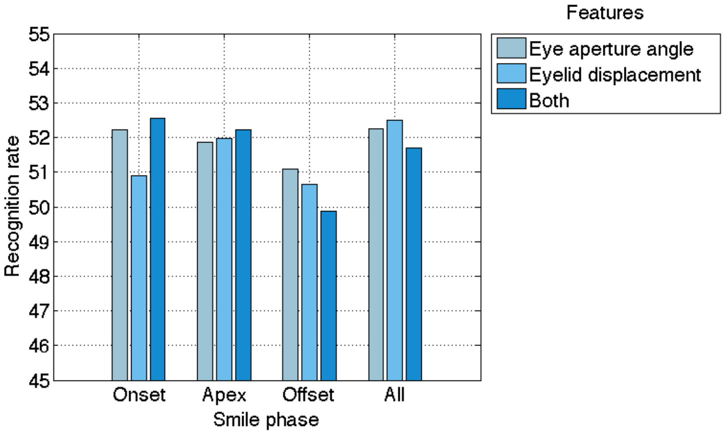

Figure 25: CHMM classification results for eye aperture angle, eyelid displacement and the two combined.

Similarly to CHMM and SVM classifiers trained on head and lip features, the SVM classifiers trained on eye features outperform the CHMM classifiers trained on eye features (see Figure 22 and 23). The results for the SVM classifiers trained on certain combinations of the eye features are shown in Figure 26 and 27. The highest accuracy for all eye features combined is 66.83%. The combination of (non-Bézier) eye aperture angle features and eyelid displacement features excedes this with an accuracy of 67.91% for the combination of all phases (see Figure 27). The highest accuracy, however, is achieved by Bézier eye aperture angle features and eyelid displacement features from all phases (68.51%), which is shown in Figure 26. Hence, SVM classifiers trained on the latter combination of eye features are selected for the fusion with SVM classifiers trained on head and lip features.

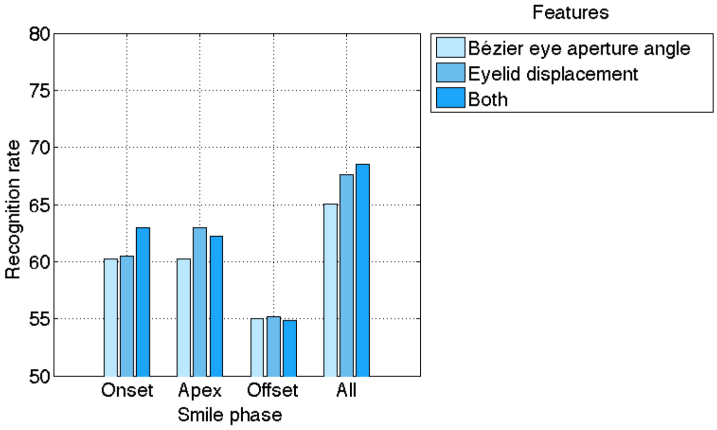

Figure 26: SVM classification results for Bézier eye aperture angle, eyelid displacement and the two combined.

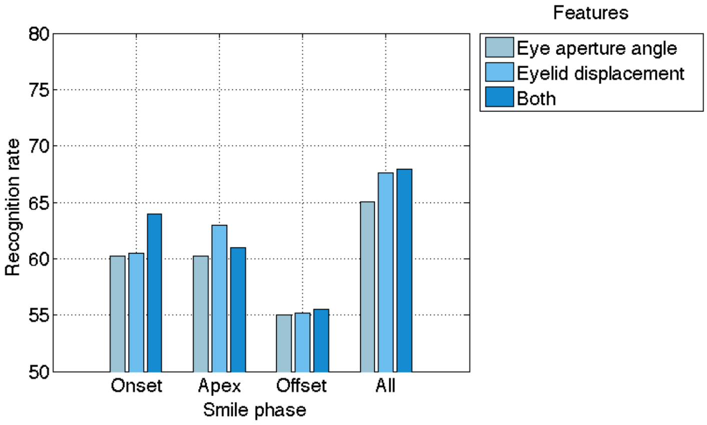

Figure 27: SVM classification results for eye aperture angle, eyelid displacement and the two combined.

When examining the SVM results of the individual eye features, it can be noticed that eyelid displacement features are most discriminating for individual phases (see Figure 23). Furthermore, comparing the results for individual phases, each eye feature is most discriminating during the apex and least discriminating during the offset.

4.3.4 The influence of fusion on classification

To investigate the influence of fusion on the classification of posed and spontaneous enjoyment smiles, classifiers trained on different features are fused and the results are compared to the classification results achieved by each of the features individually. As previously mentioned in Section 3.5, late fusion is used for the CHMM classifiers and early fusion is used for the SVM classifiers.

The classification results for the fusion of CHMM classifiers trained on head and lip features are shown in Figure 28. As can be seen, the fusion of head and lip features only improves classification when combining all phases (54.21%). For individual phases, lip features are more discriminating than the fusion of head and lip features.

Furthermore, the classification results for the fusion of CHMM classifiers trained on head and eye features are shown in Figure 29. The fusion of head and eye features improves classification for the onset and offset, but is outperformed by eye features during the apex and for all phases combined. The highest head-eye fusion accuracy is achieved when combining all phases (52.98%), but this is exceeded by eye features of all phases combined (53.13%).

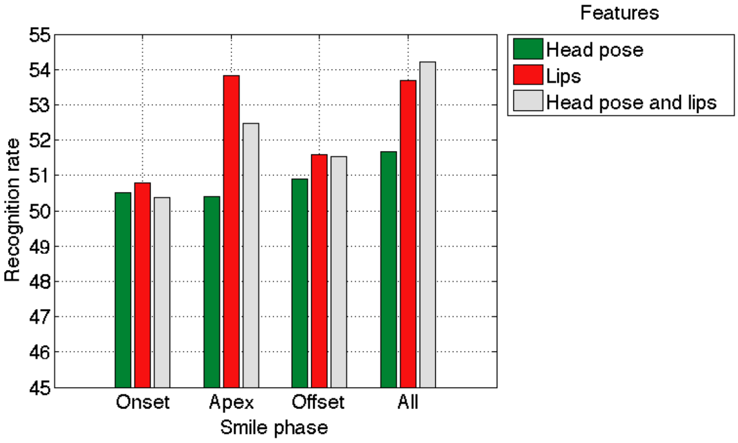

Figure 28: CHMM classification results for the fusion of head pose and lip features.

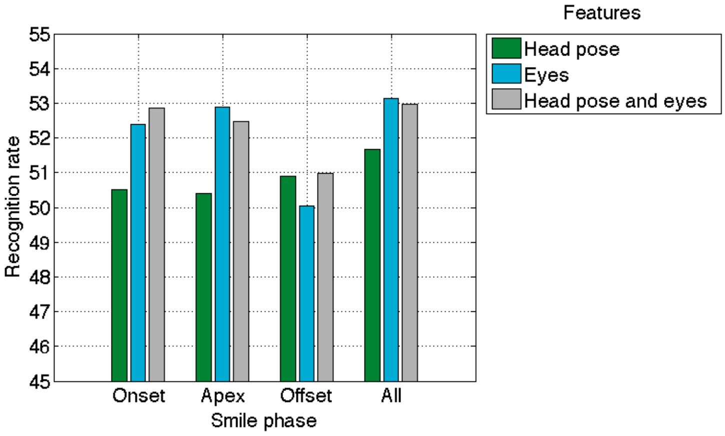

Figure 29: CHMM classification results for the fusion of head pose and eye features.

Moreover, the classification results for the fusion of CHMM classifiers trained on lip and eye features can be seen in Figure 30. The fusion of lip and eye features does not improve classification for any of the phases.

Finally, the classification results for the fusion of CHMM classifiers trained on head, lip and eye features are shown in Figure 31. The fusion of the features from the three regions improves classification merely for the onset and apex. The highest accuracy is reached by the fusion of all features during the apex (54.14%).

When comparing the results for the fusion of every combination of features in Figure 36, it can be noticed that the fusion of head and lip features achieves the highest accuracy (54.21%), followed by the fusion of all features (54.14%), the fusion of lip and eye features (53.46%) and, finally, the fusion of head and eye features (52.98%).

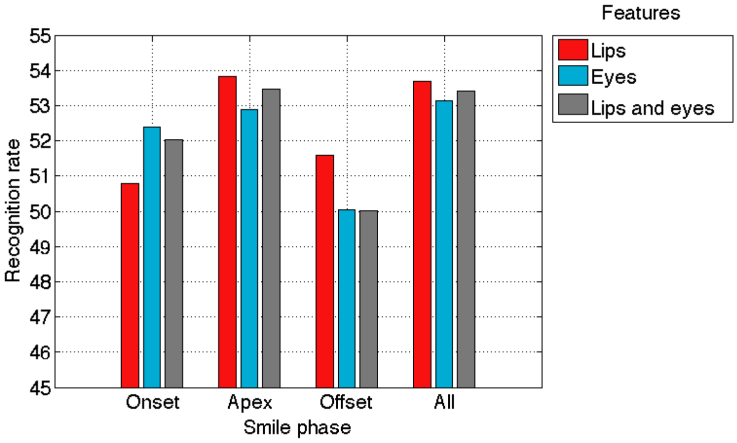

Figure 30: CHMM classification results for the fusion of lip and eye features.

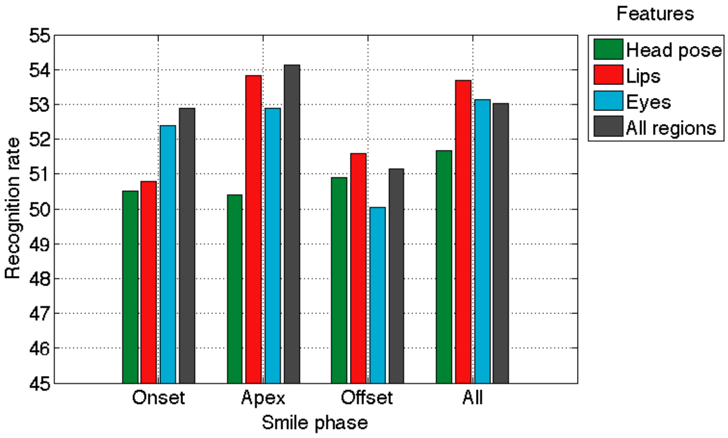

Figure 31: CHMM classification results for the fusion of all features.

The classification results for SVM classifiers trained on the fusion, or concatenation, of head and lip features are shown in 32. Similarly to the corresponding CHMM classifiers, the fusion of head and lip features is generally outperformed by the lip features. The highest accuracy for the fusion of head and lip features is achieved for all phases combined (72.87%), but this is exceeded by the lip features for all phases combined (73.90%).

Furthermore, the classification results for SVM classifiers trained on the fusion of head and eye features are shown in Figure 33. The fusion of head and eye features improves accuracy during the apex and for all phases combined, but is outperformed by head features during the onset and offset. The highest accuracy for the fusion of head and eye features is achieved when combining all phases (72.46%) and this exceeds all head and eye feature results.

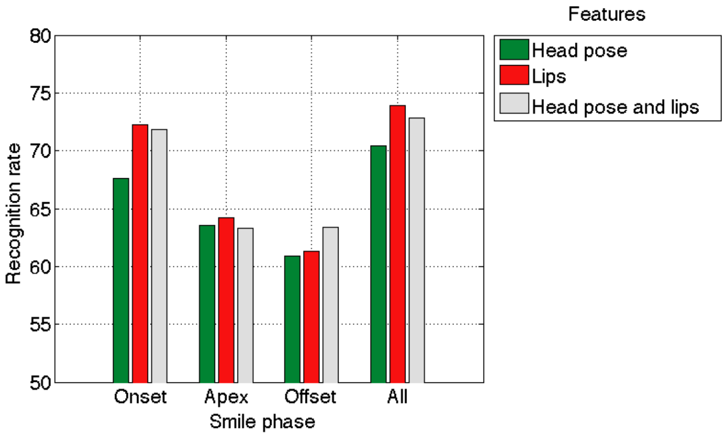

Figure 32: SVM classification results for the fusion of head pose and lip features.

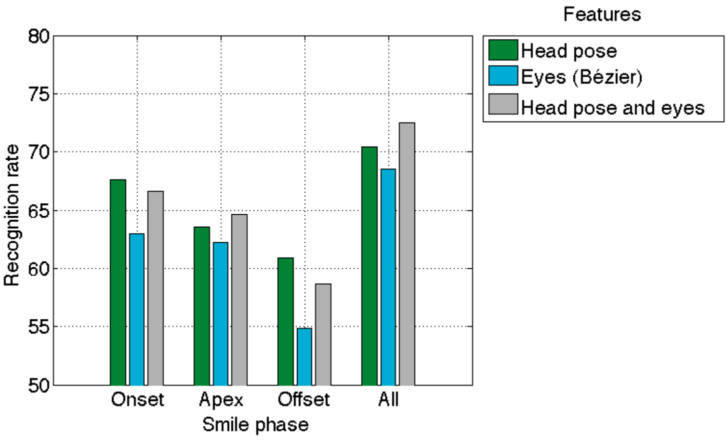

Figure 33: SVM classification results for the fusion of head pose and eye features.

Moreover, the classification results for SVM classifiers trained on the fusion of lip and eye features are shown in Figure 34. The fusion of lip and eye features improves classification for all phases, except for the onset. The highest accuracy is reached by fusing lip and eye features from all phases (74.60%).

Finally, the classification results for the SVM classifiers trained on all features are shown in Figure 35. The fusion of all features improves classification for all phases, except for the onset. The highest accuracy is reached by training on all features from all phases (76.06%).

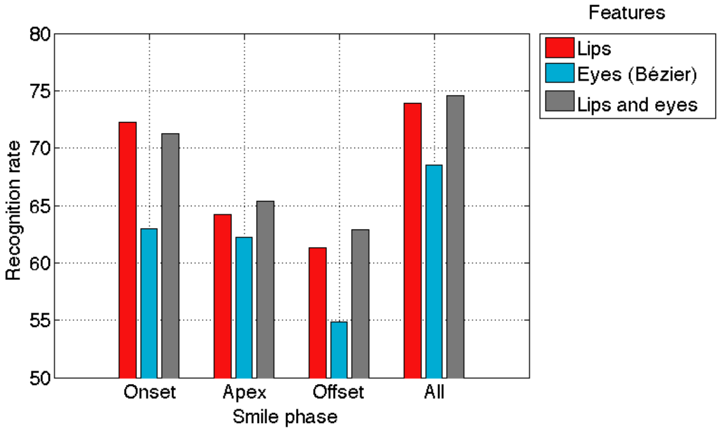

Figure 34: SVM classification results for the fusion of lip and eye features.

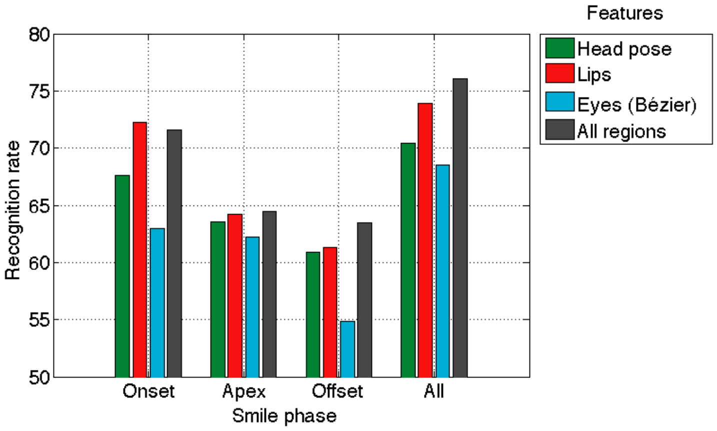

Figure 35: SVM classification results for the fusion of all features.

When comparing the fusion combinations in Figure 37, it can be noticed that for individual phases, the fusion of head pose and lips achieves the highest accuracy during the onset (71.81%). Furthermore, the highest accuracy is reached by fusing features from all regions and all phases (76.06%).

Moreover, the effect of the additional duration feature is also shown in Figure 37. The addition of the duration feature results in a better classification for every phase, in particular for the onset and apex phases. Finally, the overall highest accuracy is reached by the fusion of all features from all phases with the duration feature (77.85%).

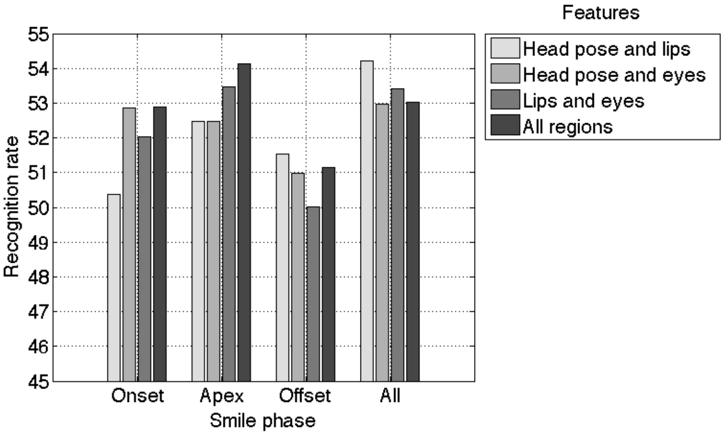

Figure 36: CHMM classification results for the fusion of certain combinations of features and all features.

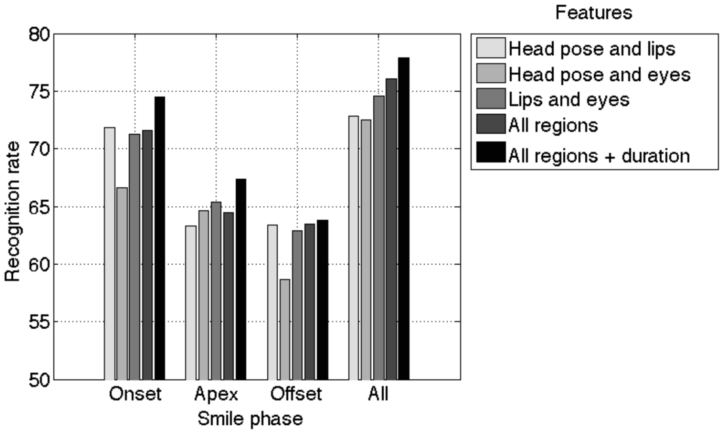

Figure 37: SVM classification results for the fusion of certain combinations of features and all features. Furthermore, the effect of an additional duration feature is shown.

4.3.5 Comparison to state-of-the-art methods

Dibeklioğlu et al. (2012) evaluated several state-of-the-art method on the UvA-NEMO Smile Database. The results of these methods and the methods of this study are shown in Table 3.

Method	Correct classification rate (%)
Dibeklioğlu et al. (2012), Mid-level fusion.	87.02
Dibeklioğlu et al. (2012), Eyelid Features.	85.73
Proposed, SVM	77.85
Cohn et al. (2004)	77.26
Pfister et al. (2011)	73.06
Dibeklioğlu et al. (2010)	71.05
Proposed, CHMM	54.21

Table 3: Correct classification rates on the UvA-NEMO Smile Database. Adopted from Dibeklioğlu et al. (2012).

Besides comparison to state-of-the-art methods that all use temporal information to some extent, the proposed methods can be compared to the classification rate achieved for the classification of posed and spontaneous smiles in still images. Such a comparison shows the influence of temporal facial information on the classification of posed and spontaneous enjoyment smiles. Zhang et al. (2011) achieved an accuracy of roughly 70% on a more conditioned dataset than the UvA-NEMO Smile Database using SIFT appearance based features and FAP geometric features from the face. Furthermore, Petridis et al. (2009) achieved an accuracy of about 75% on a more conditioned dataset, fusing information from the face and the head. The classification of posed and spontaneous enjoyment smiles in still images is not performed on the UvA-NEMO Smile Database, but the resulting classification rate is likely lower than 75% or 70% because the UvA-NEMO Smile Database is less conditioned.

5 Discussion

In the conducted experiments, the SVM classifiers outperform the CHMM classifiers greatly. The highest accuracy for CHMM classifiers is achieved by fusing head and lip features from all phases (54.21%) and the corresponding SVM accuracy is 72.87%. This low performance of the CHMM classifiers is most likely due to their optimization and not due to the use of a CHMM per se. This is because the CHMM classifier of Dibeklioğlu et al. (2010) evaluated on the UvA-NEMO Smile Database reaches an accuracy of 71.05%, using features from the eyes only (see Table 3). Although the eye features used in this study slightly differ from the eye features used by Dibeklioğlu et al. (2010)[6], this cannot be the reason of the huge difference in performance.

Because the CHMM classification results are low and lie within a small range (roughly 50% to 55%), they are not very reliable and, therefore, not very useful for analysis. The rest of this section therefore focuses on the SVM classification results rather than on the CHMM classification results.

In the conducted experiments, head pitch is found to be the most discriminating head pose feature. This finding could provide some basis for investigating the suggestion of Cohn et al. (2004) that head pitch can have a positive correlation with smile intensity, both movements embedded within a coordinated motor structure. This suggestion has arisen from their finding that in spontaneous smiles of embarrassment, head pitch is negatively correlated with smile intensity, which means that the head moves downwards when the smile reaches the apex and moves upwards during the offset. This is quite typical of smiles of embarrassment. It is plausible that indeed an opposite pattern can be seen in smiles of joy: head pitch and smile intensity increasing and decreasing together. In addition, it is plausible that this correlation is not seen in posed smiles because coordinated motor structures are possibly not employed when posing a smile.

Furthermore, lip features are the most discriminating features for individual and all phases, followed by features from the head and, lastly, features from the eyes. The onset lip angle features outperform the lip angle features during the apex and offset. This result is consistent with the findings of Schmidt et al. (2006), which indicate that spontaneous and posed smiles differ in lip corner speed speed and amplitude during the onset. In spontaneous smiles the onset speed is slower and the amplitude smaller than in posed smiles. However, during the offset the differences in speed and amplitude are insignificant. This explains the higher accuracy of onset lip angle features. The seemingly contradictory low performance of the lip displacement features during the onset can be explained by the fact that, in constrast to the lip angle features, no speed features were computed for lip displacement. This feature is therefore less discriminating than the lip angle features.

Another important finding is that the eye features used in this study are not very discriminating in comparison with the eye features used by Dibeklioğlu et al. (2012), which achieve an accuracy of 85.73%. More specifically, the eye aperture angle features are the least discriminating features for every phase. This is probably because these eye aperture angle features define the angles on the inner corners of the eyes, as opposed to the outer eye corner angle features used by Dibeklioğlu et al. (2012). When the *orbicularis oculi, pars lateralis* muscle contracts[7], the outer corners of the eyes are likely to be more narrowed than the inner corners of the eyes and, therefore, be more informative than the inner corners of the eyes. Orbicularis oculi activity is important according to Ekman and Friesen (1982) because it is present during spontaneous smiles, but not during posed smiles. Although this finding is challenged by findings from Krumhuber and Manstead (2009), which indicated that the orbicularis oculi is approximately as frequently present in posed smiles as in spontaneous smiles, it could explain the present finding that the eye aperture angle features are less discriminating than the eyelid displacement

[6] The eye aperture angles in this study are computed for the inner eye corners. In contrast, the eye aperture angles used by Dibeklioğlu et al. (2010) are computed for the outer eye corners. The eyelid displacement features are identical in both studies.

[7] The orbicularis, pars lateralis muscle raises the cheeks, narrows the opening of the eyes and forms wrinkles around the eyes.

features. This is because the landmarks used to compute the eyelid displacement features are closer in distance to the outer corner of the eyes and, thus, eyelid displacement might reflect orbicularis oculi activity better than the eye aperture angle features. This could explain the higher accuracy of eyelid displacement, in particular during the apex, when (if present) the orbicularis oculi is contracted most.

The fusion of all features leads to a better classification than each of the features on their own, except during the onset, where lip features are slightly more descriptive. The addition of even more temporal information, achieved by fusing all features with the additional duration feature, improves classification for every phase, in particular for the onset and apex phases. This finding is in accordance with findings from Ekman and Friesen (1982), which indicated that posed and spontaneous smiles differ in duration during the onset and apex phases. In false (posed) smiles, apex duration is usually too long and the onset duration too short, which gives an abrupt appearance to the smile. In addition, the offset often appears not smooth in false smiles, however, no difference in duration is provided by Ekman and Friesen (1982) for this phase. This could indicate that there is a minimal or insignificant difference in duration during the offset phases of posed and spontaneous smiles, which is reflected in this study where the duration feature has a minimal effect on classification during the offset.

Furthermore, for individual phases, the fusion of certain combinations of features only occasionally leads to a better classification than the features on their own. However, when onset, apex and offset phases are fused, the fusion of every combination of features leads to a better classification than each of the features on their own, except for the fusion of head and lip features. In fact, the fusion of the onset, apex and offset phases generally leads to a better classification than the phases on their own. This indicates that the use of more temporal information generally improves classification.

When comparing the highest recognition rate achieved in the present study (77.85%) with state-of-the-art methods in Table 3, it can be seen that the performance of the proposed classifier is quite good. It outperforms three state-of-the-art methods. Furthermore, the influence of temporal facial information on the classification of posed and spontaneous enjoyment smiles is shown when comparing the proposed classifier to classification of posed and spontaneous enjoyment smiles in still images. Whereas classification in still images achieves a recognition rate of roughly 75%, classification using temporal facial information achieves a recognition rate of 77.85% in the present study. This shows that the use of temporal information improves the classification of posed and spontaneous enjoyment smiles with a recognition rate improvement of at least 4%[8].

[8] At least 4% because the UvA-NEMO Smile Database is less conditioned. This means that it is likely that the still image classification methods being compared with the proposed method of this study perform worse on the UvA-NEMO Smile Database than on the dataset on which they were evaluated. Therefore, the use of temporal information might improve the classification of posed and spontaneous enjoyment smiles more than shown in this comparison.

6 Conclusion

The influence of temporal information on the classification of posed and spontaneous enjoyment smiles is investigated. CHMM and SVM classifiers are trained on features from the head, lips and eyes to investigate the influence of these features on the classification of posed and spontaneous enjoyment smiles. In addition, the influence of the fusion of classifiers trained on different features on classification is investigated.

The results indicate the head motion positively influences the classification of posed and spontaneous enjoyment smiles, especially head pitch is a discriminating feature. Furthermore, lip corner movement has the greatest influence on classification. Lip corner movement is more discriminating than head motion and eyelid movement. In addition, eyelid movement has the least positive influence on classification in this study. The results for fusion indicate that for individual phases, the fusion of all features generally leads to a better classification than using each of the features on their own. However, the fusion of certain combinations of features only occasionally leads to a better classification for individual phases. Nevertheless, when onset, apex and offset phases are fused, the fusion of certain combinations of features generally leads to a better classification than each of the features on their own.

Finally, the comparison between the performance of classification of posed and spontaneous enjoyment smiles in still images and in moving images, as is performed in the present study, shows that the use of temporal facial information improves the classification of posed and spontaneous enjoyment smiles.

7 Future work

The CHMM classifers used in the present study have a low performance. A better optimization of these classifiers could improve results greatly. With better and more reliable results, the influence of individual features and the fusion of certain combinations of features for CHMM classifiers can be compared to the corresponding results of SVM classifiers.

Furthermore, classification using SVM classifiers with a linear kernel and parameter selection should be investigated. In this study, the focus is on SVM classifiers with RBF (with parameter selection) because these are found to have a better performance when trained on individual features. However, SVM classifiers with a linear kernel without parameter selection perform slightly better when trained on the fusion of all features. Therefore, SVM classifiers with a linear kernel and parameter selection could possibly further improve classification.

Finally, head pitch is found to be the most discriminative head pose feature, however, further investigation is required to discover the discriminative value of head pitch in posed and spontaneous smiles.

References

Jeffrey F Cohn and Karen L Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02):121–132, 2004.

Jeffrey F Cohn, Lawrence Ian Reed, Tsuyoshi Moriyama, Jing Xiao, Karen Schmidt, and Zara Ambadar. Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 129–135. IEEE, 2004.

Marco Del Giudice and Livia Colle. Differences between children and adults in the recognition of enjoyment smiles. *Developmental psychology*, 43(3):796, 2007.

Hamdi Dibeklioğlu, Roberto Valenti, Albert Ali Salah, and Theo Gevers. Eyes do not lie: spontaneous versus posed smiles. In *Proceedings of the international conference on Multimedia*, pages 703–706. ACM, 2010.

Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *Computer Vision–ECCV 2012*, pages 525–538. Springer, 2012.

P. Ekman and W.V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*, chapter 2, pages 25–42. WW Norton & Company, 2009.

Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6 (4):238–252, 1982.

Carl-Herman Hjortsjö. *Man's face and mimic language*. Studentlitteratur, 1969.

Dacher Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3):441, 1995.

Eva Krumhuber and Arvid Kappas. Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior*, 29(1):3–24, 2005.

Eva G Krumhuber and Antony SR Manstead. Can duchenne smiles be feigned? new evidence on felt and false smiles. *Emotion*, 9(6):807, 2009.

Jenn-Jier James Lien. *AUTOMATIC RECOGNITION OF FACIAL EXPRESSIONS USING HIDDEN MARKOV MODELS AND ESTIMATION OF EXPRSSION INTENSITY*. PhD thesis, Washington University, 1998.

Stavros Petridis, Hatice Gunes, Sebastian Kaltwang, and Maja Pantic. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 23–30. ACM, 2009.

Tomas Pfister, Xiaobai Li, Guoying Zhao, and M Pietikainen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 868–875. IEEE, 2011.

Rosalind W Picard. Toward machines with emotional intelligence. In *ICINCO (Invited Speakers)*, pages 29–30. Citeseer, 2004.

Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, 30(1):37–52, 2006.

Karen L Schmidt, Sharika Bhattacharya, and Rachel Denlinger. Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Journal of nonverbal behavior*, 33(1):35–45, 2009.

Caifeng Shan, Shaogang Gong, and Peter W McOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *BMVC*, pages 297–306, 2006.

Michel F Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

Paul F Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, 75(371):609–615, 1980.

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Geometry vs. appearance for discriminating between posed and spontaneous emotions. In *Neural Information Processing*, pages 431–440. Springer, 2011.

Appendices

A Additional head pose results

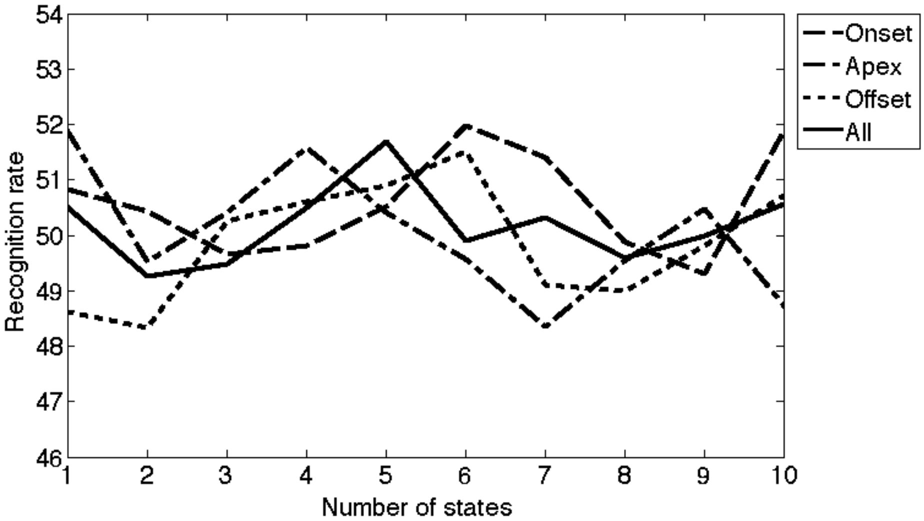

Figure A.1: Recognition rates for CHMMs trained on head pose with different numbers of states.

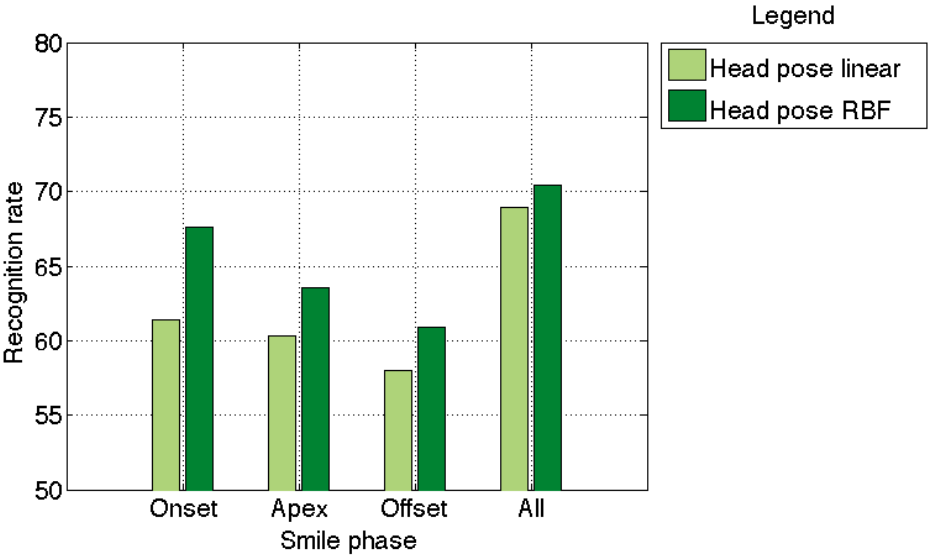

Figure A.2: Recognition rates for SVMs trained on head pose with RBF or a linear kernel.

B Additional lip feature results

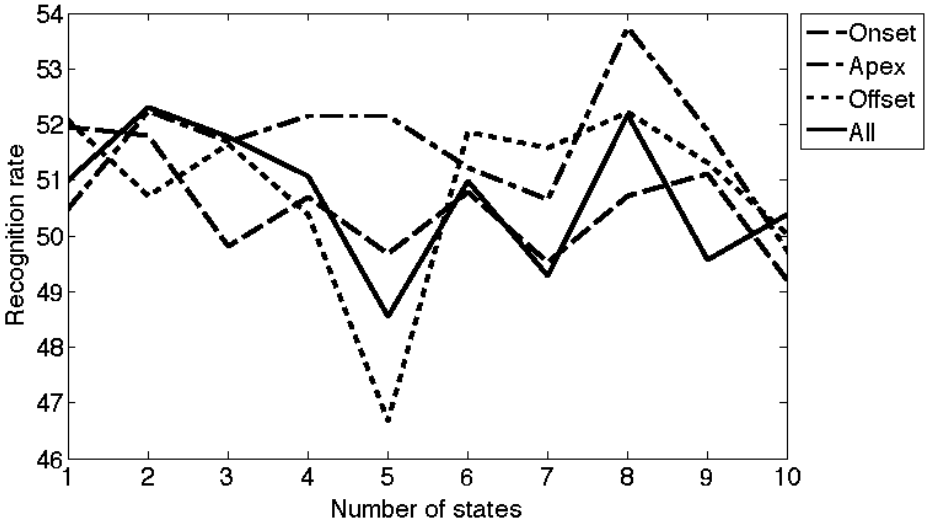

Figure B.1: Recognition rates for CHMMs trained on lip displacement features with different numbers of states.

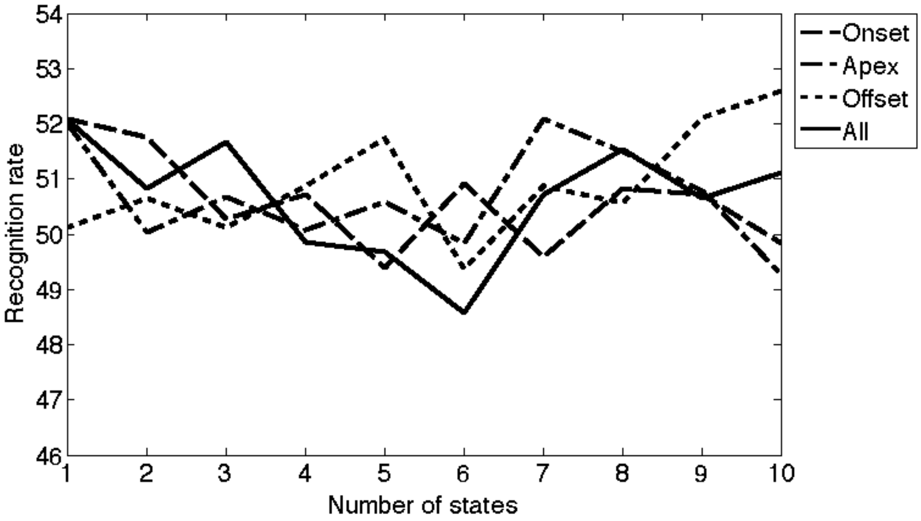

Figure B.2: Recognition rates for CHMMs trained on lip angle features with different numbers of states.

C Additional eye feature results

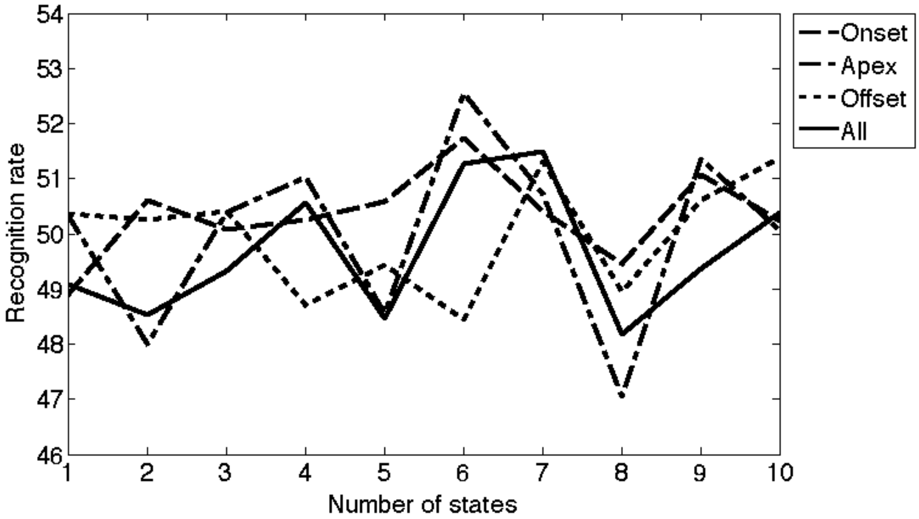

Figure C.1: Recognition rates for CHMMs trained on Bézier eye aperture angle features with different numbers of states.

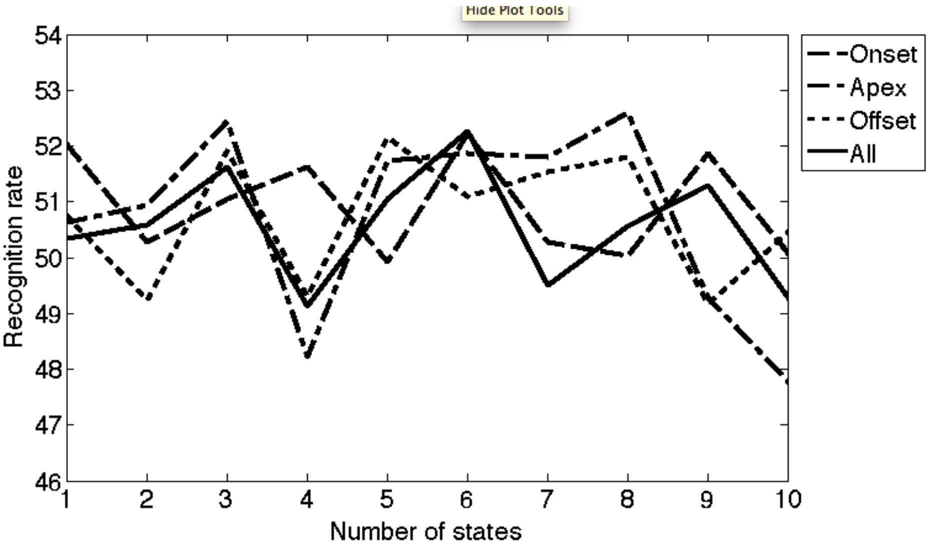

Figure C.2: Recognition rates for CHMMs trained on (non-Bézier) eye aperture angle features with different numbers of states.

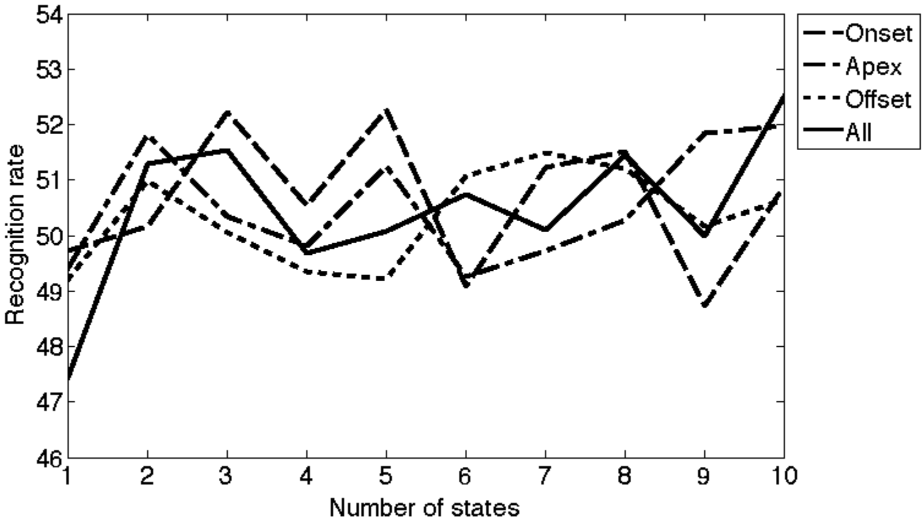

Figure C.3: Recognition rates for CHMMs trained on eyelid displacement features with different numbers of states.

D Additional fusion results

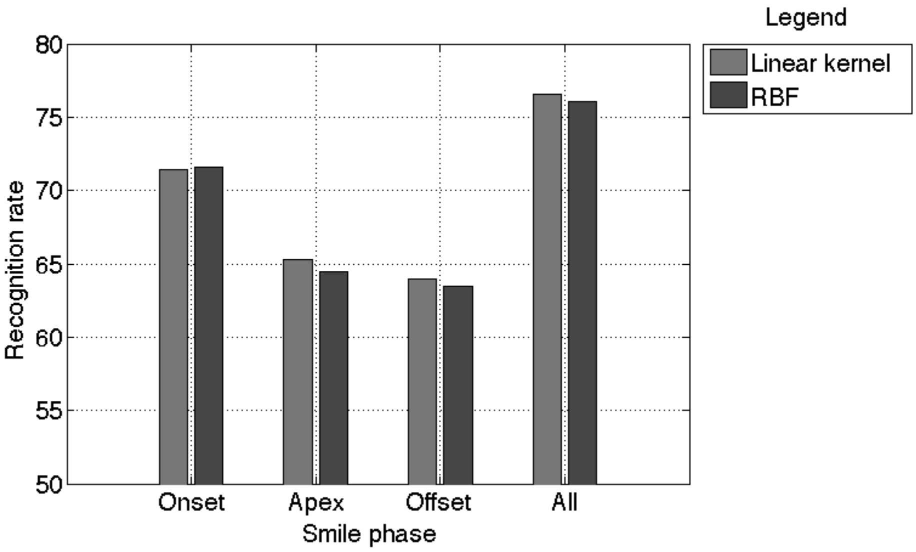

Figure D.1: Classification results for SVMs trained on all features with RBF or a linear kernel.