# Static vs. Dynamic Modeling of Human Nonverbal Behavior from Multiple Cues and Modalities

Stavros Petridis
Dep. of Computing
Imperial College London
UK
sp104@doc.ic.ac.uk

Hatice Gunes
Dep. of Computing
Imperial College London
UK
hgunes@doc.ic.ac.uk

Sebastian Kaltwang
Dep. of Computer Science
Univ. of Karlsruhe
Germany
s.kaltwang@gmail.com

Maja Pantic
Dep. of Computing
Imperial College London, UK
EEMCS, Univ. Twente, NL
m.pantic@imperial.ac.uk

## ABSTRACT

Human nonverbal behavior recognition from multiple cues and modalities has attracted a lot of interest in recent years. Despite the interest, many research questions, including the type of feature representation, choice of static vs. dynamic classification schemes, the number and type of cues or modalities to use, and the optimal way of fusing these, remain open research questions. This paper compares frame-based vs. window-based feature representation and employs static vs. dynamic classification schemes for two distinct problems in the field of automatic human nonverbal behavior analysis: multicue discrimination between posed and spontaneous smiles from facial expressions, head and shoulder movements, and audio-visual discrimination between laughter and speech. Single cue and single modality results are compared to multicue and multimodal results by employing Neural Networks, Hidden Markov Models (HMMs), and 2- and 3-chain coupled HMMs. Subject independent experimental evaluation shows that: 1) both for static and dynamic classification, fusing data coming from multiple cues and modalities proves useful to the overall task of recognition, 2) the type of feature representation appears to have a direct impact on the classification performance, and 3) static classification is comparable to dynamic classification both for multicue discrimination between posed and spontaneous smiles, and audio-visual discrimination between laughter and speech.

## Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.m [**Computer Applications**]: Miscellaneous

*ICMI-MLMI'09,* November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11 ...$10.00.

## General Terms

Multi-modal fusion, assimilation and processing

## Keywords

static classification, dynamic classification, frame-based representation, window-based representation, multicue and multimodal fusion.

## 1. INTRODUCTION

In the day-to-day world humans naturally combine multiple channels and modalities to communicate with others [27]. Thus, human nonverbal behavior can be recognized from a broad range of behavioral cues like facial expressions, head and hand gestures and non-linguistic vocalizations [7], [27]. Despite the available range of cues and modalities in human-human interaction, past research on affect/behavior sensing and recognition has mainly focused on single modalities like facial expressions or audio, and on data that has been acted on demand or acquired in laboratory settings [27]. Automatic systems using multiple cues and modalities, and capable of handling spontaneous data acquired in naturalistic settings have only recently emerged [7]. This in turn triggered many other research questions: what features to extract [24], which classification schemes to employ, which cues or modalities to use, and how to combine them [7].

From the automatic sensing perspective, human nonverbal behavior analysis can be performed by either using the features from one frame at a time, or by considering the sequential nature of the frame sequence as in a time series. In the literature, these two approaches are referred to as *static or frame-based* and *dynamic or sequence-based* classification, respectively [6], [25]. Commonly used static classifiers are Support Vector Machines (SVM), Neural Networks (NN) and decision trees (C4.5). Hidden Markov Models(HMM) and their variations (e.g., Coupled Hidden Markov Models (CHMM)) constitute the well known dynamic classifiers.

As early work on automatic emotion recognition has mostly focused onto a simplified problem of recognition of small number of classes of posed (deliberately displayed), facial expression images, for many years static classification has been the trend in the automatic affect recognition field [27].

However, as the research field has shifted its focus from static images to multicue and multimodal analysis, and from acted data to spontaneous and naturalistic data, more recent approaches have been exploring the use of dynamic classification techniques to the aim of improving recognition accuracy (e.g., [3], [6], [28]).

Human nonverbal behavior is inherently continuous and sequential. It consist of streams of multicue (e.g., smiling accompanied by downward head pitch) and multimodal (e.g., smiling accompanied by a high acoustic pitch) data. Therefore, machine learning methods like Dynamic Bayesian Networks (DBNs) are considered to be better suited for spontaneous affective behavior recognition. They are known to well model the temporal activity incorporated within sequential affect data (e.g., [28]).

For emotional speech recognition, either global statistics features are calculated and fed to a static classifier or short-term features are commonly used for dynamic modeling via HMMs [25]. The researchers claim that in the static classification case dynamic properties of human affective behavior should be captured by the features, while in the latter case, they are dealt with by the classifier.

Some researchers reported reported that dynamic classifiers are better suited for person dependent facial expression recognition (e.g., [3]). This was attributed to the fact that dynamic classifiers are more sensitive to both differences in terms of appearance change and differences in temporal patterns among individuals. Static classifiers instead were reported as being more reliable when the the frames represent the apex of an expression [3]. Other researchers reported that the frame-based classification outperforms the sequence-based classification in the task of temporal segment detection from face and body display (e.g., [6]). It was shown that the accuracy of affect recognition increases significantly if static classifiers are employed using the apex frames from face and body display compared to that of feeding a whole sequence to a dynamic classifier such as an HMM [6].

The main challenge faced when comparing static and dynamic classification relates to the utilised features. Speech-based emotion recognition has mostly used turn-wise statistics of acoustic features followed by a static classification or frame-level features followed by a dynamic classification [23]. Vogt et al. [25] argue, that as such works use different feature representation for static and dynamic classification, it is not possible to clearly attribute the higher recognition accuracy to either classification technique (dynamic vs. static) [25]. It could well be the case that the utilised features derived cause the difference in the obtained recognition accuracies. For emotional speech recognition, Vogt et al. claim that static classification performs better as more feature types can be exploited (e.g., suprasegmental acoustic features like jitter or shimmer to measure voice quality) [25]. However, if the same feature types are used (e.g., only MFCCs and energy), HMMs appear to outperform static modeling techniques.

Despite the aforementioned efforts, to the best of our knowledge, no systematic evaluation of static and dynamic classification methods for automatic human affective behavior analysis from multicue and multimodal data has been reported to date. To this aim, in this paper, we focus on two facets of human nonverbal behavior sensing and recognition: automatically discriminating between posed and spontaneous smiles from multiple visual cues [22], and automatically detecting laughter vs. speech from audio and visual modalities [16].We use both spontaneous and posed (as opposed to posed only) displays of smiles from the MMI facial expression database [12], and spontaneous laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [9]. We focus on person-independent recognition which makes the task of human nonverbal behavior analysis even more challenging. We provide details of the single cue or single modality recognition merely in order to obtain a term of reference for the performance of the multicue and multimodal recognition. We then evaluate static vs. dynamic classification by employing Neural Networks and (coupled) Hidden Markov Models for the two problems at hand. The experimental results obtained show the following: 1) for both static and dynamic classification, fusing data coming from multiple cues and modalities proves useful to the overall task of recognition, 2) the type of feature representation appears to have a direct impact on the classification performance, and 3) static classification is comparable to dynamic classification both for multicue discrimination between posed and spontaneous smiles, and audio-visual discrimination between laughter and speech.

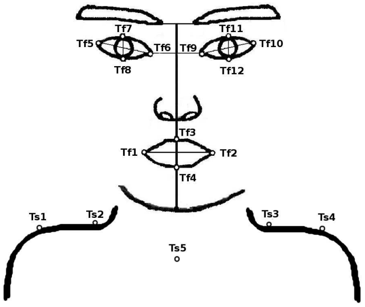

**Figure 1: Tracked points $T_{f1} \ldots T_{f12}$ of the face and tracked points $T_{s1} \ldots T_{s5}$ of the shoulders.**

In the remainder of the paper, the problem of distinguishing between acted and spontaneous smiles is referred to as Case Study 1, and the problem of discriminating between laughter and speech is referred to as Case Study 2.

## 2. DATABASES

### 2.1 Case Study 1

For Case Study 1, we used 201 videos displaying acted and spontaneous smiles from 43 subjects from the MMI facial expression database [12]. The MMI facial expression database has two parts: a part containing deliberately displayed facial expressions and a part containing spontaneous facial displays. The acted part contains videos depicting facial expressions of single Action Unit (AU) activation (e.g., AU12 or AU13), multiple AU activations (e.g., AU6 and AU12), and six basic emotions. The spontaneous part of the database contains videos of spontaneous facial displays. We used 99 videos from the acted part and 102 from the spontaneous part. The recordings of the spontaneous part were made partially in a TV studio, using a uniform background and constant lighting conditions, and partially in subjects' usual environments (e.g., home), where they were shown segments from comedies, horror movies, and fear-factor series. These recordings contain mostly facial expressions of different kinds of laughter, surprise, and disgust expressions, which were accompanied by (often large) head motions. We selected the videos that contain facial expressions of different kinds of smiles (AU12 or AU13).

## 2.2 Case Study 2

For Case Study 2 we used the AMI Meeting Corpus which consists of 100 hours of meetings recordings where people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face (720 x 576 pixels, 25 frames per second) and the related individual headset audio recordings (16 kHz). The language used in the meetings is English and the speakers are mostly non-native speakers. For our experiments we used seven meetings (IB4001 to IB4011) and the relevant recordings of eight participants (6 young males and 2 young females) of Caucasian origin with or without glasses and no facial hair. All laughter and speech segments were pre-segmented based on audio. Initially, laughter segments were selected based on the annotations provided with the AMI Corpus. After examining the extracted laughter segments we only kept those that do not co-occur with speech and where the laughter is clearly audible. Speech segments were also determined by the annotations provided with the AMI Corpus. We selected those that do not contain long pauses between two consecutive words. In total, we used 114 audio-visual laughter segments and 92 audio-visual speech segments.

# 3. FEATURE EXTRACTION

## 3.1 Case Study 1

In order to distinguish between posed and spontaneous smiles based on multiple visual cues, we track facial feature points, head and shoulder movements.

**Head features (He).** To capture the head motion we employ the Cylindrical Head Tracker developed by Xiao et al. [26]. The head tracker estimates the six degrees of freedom of head motion: horizontal and vertical position in the scene, distance to the camera (i.e. scale), pitch, yaw and roll. This is denoted as the set of parameters $T_h = \{T_{h1} \ldots T_{h6}\}$ with dimensions $n*6$. Here $n$ is the number of frames of an input sequence.

**Facial expression features (Fa).** To capture the facial motion displayed during a smile we track 12 facial points, points $T_{f1} - T_{f4}$ and points $T_{f5} - T_{f12}$, as illustrated in Fig. 1. These points are the corners (extremities) of the eyes (8 points) and the mouth (4 points). To track these facial points we used the Patras - Pantic particle filtering tracking scheme [13]. For each video segment containing $n$ frames, we obtain a set of $n$ vectors containing 2D coordinates of the 12 points tracked in $n$ frames ($T_f = \{T_{f1} \ldots T_{f12}\}$ with dimensions $n * 12 * 2$).

**Shoulder features (Sh).** 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. 1). The stable point is used to remove any rigid motion of the torso. We use the standard Auxiliary Particle Filtering (APF) [17] to this aim. This scheme is less complex and faster compared to the Patras - Pantic particle filtering tracking scheme, 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 $T_s = \{T_{s1} \ldots T_{s5}\}$ with dimensions of $n * 5 * 2$.

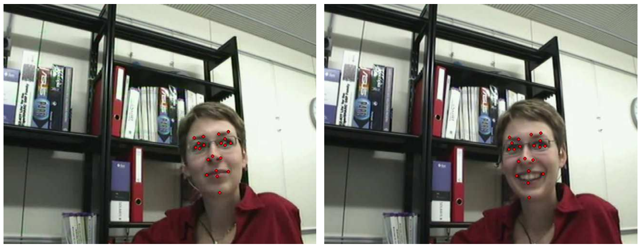
(a) Frame 2 (b) Frame 122

**Figure 2: Example of a laughter episode, from the AMI corpus, with illustrated facial point tracking results.**

The final feature set obtained for each visual cue contains tracked points as well as distances, angles and speed and is described in detail in Table 1. Further details of the tracking and feature extraction procedure can be found in [22].

After some preliminary experiments, we chose to use $Fa[21 : 40]$, $He[7 : 12]$, and $Sh[5 : 8]$ for this study. These features are derived from the previously calculated features and are normalized with respect to the neutral frame.

## 3.2 Case Study 2

In order to discriminate between laughter and speech, information is extracted simultaneously from the audio and visual channels as follows.

**Spectral features (Sp).** Spectral or cepstral features, such as MFCCs, have been widely used in speech recognition [18] and have also been successfully used for laughter detection [8]. Only the first 6 MFCCs are used, given the findings in [8], which are computed every 20ms over a window of 40ms. It has been shown that both 50 frames per second (fps) and 100 fps have the same performance for the task at hand [14] so we chose to use 50 fps.

**Prosodic features (PE).** The two most commonly used prosodic features in studies on emotion detection are pitch and energy [27]. Pitch is the perceived fundamental frequency of a sound. While the actual fundamental frequency can be precisely determined through physical measurement, it may differ from the perceived pitch. Bachorowski et al. [1] found that the mean pitch in both male and female laughter was higher than in modal speech. Pitch was computed in each frame using the same algorithm as Praat [2]. The energy feature used is the Root-Mean-Square (RMS) energy. Those features are extracted in the same frame rate as the MFCC coefficients, i.e., every 20ms over a window of 40ms.

**Visual features (Fa).** Changes in facial expression are captured by tracking 20 facial points as shown in Fig. 2. These points are the corners (extremities) of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points) and the chin (1 point). Tracking was done using the tracker proposed in [13]. For each video segment containing n frames, we obtain a set of n vectors containing 2D coordinates of the 20 points tracked in n frames. Using a Point Distribution Model (PDM), head movement is decoupled from facial expression. Using the approach proposed in [5], we extract 5 features per frame, which encode the

**Table 1: Description of the feature set used in case study 1 and obtained from tracking each face (Fa), head (He) and shoulders (Sh) cues.**

| | |
|---|---|
| Fa[1:8] | x- and y-position of mouth points $T_{f1}$, $T_{f2}$, $T_{f3}$, $T_{f4}$ |
| Fa[9:14] | Euclidian distances between pairs of points ($T_{f1}$ and $T_{f2}$, $T_{f1}$ and $T_{f3}$) |
| Fa[15:20] | angles (between the line connecting two facial points and the y=0 line) |
| Fa[21:40] | difference of F[1:20] at time t with respect to the neutral frame |
| He[1] | x-displacement wrt neutral frame |
| He[2] | y-displacement wrt neutral frame |
| He[3] | z-displacement wrt neutral frame (zoom) |
| He[4] | roll |
| He[5] | yaw |
| He[6] | pitch |
| He[7:12] | difference of He[1:6] at time t with respect to the neutral frame |
| Sh[1] | angle of the line connecting points on the right shoulder ($T_{s1}$, $T_{s2}$) and the line y=0 |
| Sh[2] | angle of the line connecting points on the left shoulder ($T_{s3}$, $T_{s4}$) |
| Sh[3] | normalized sum of y-displacement of right shoulder points |
| Sh[4] | normalized sum of y-displacement of left shoulder |
| Sh[5:8] | difference of Sh[1:4] with respect to the neutral frame |

facial expression movements. Further details of the feature extraction procedure can be found in [15, 16].

# 4. CLASSIFICATION AND FUSION

## 4.1 Classifiers

We employ a Neural Network classifier for static classification as they are able to learn a non-linear function from examples. As dynamic classifier, we have employed HMM [19] and its variations as they have been commonly used in the literature to the aim of affect recognition from visual or audio modalities (e.g., [27]).

Affective human behavior is continuous and multi-dimensional. Therefore, the output of an HMM cannot be a discrete probability variable, a mixture of continuous variables with Gaussian distribution is used instead. Our main goal is to model a set of given (training) sequences with high accuracy. Data depicting several cues or modalities can be modeled by using several HMMs. In this case, the data streams are assumed to be independent from each other.

A Coupled Hidden Markov Model (CHMM) is a series of parallel HMM chains coupled through cross-time and cross-chain conditional probabilities. Therefore, CHMMs enable better modeling of intrinsic temporal correlations between multiple cues and modalities, and allow for true interactions between different feature sets corresponding to the same nonverbal display. In the HMM model, the probability of the next state of a sequence depends on the current state of the HMM. In the CHMM model the probability of the next state of a sequence depends on the current states of *all* HMMs.

A DBN can be used to model a CHMM [10], as illustrated in Fig. 4 and 5. The inner state is represented by a node $S$ and the output variable is represented by node $O$. $O$ depends only on the $S$ of the same timeslice and $S_{(t)}$ depends of $S_{(t-1)}$. Prior distribution for $S$ is 0, except for the start state where it is 1. The conditional probability distribution (CPD) for $(S_{(t)}, O_{(t)})$ is defined by the probabilities of the output variable and the CPD for $(S_{(t-1)}, S_{(t)})$ is defined by the transition probabilities. Fig. 4 and 5 show a DBN modeling a 3-chain CHMM. See [10] for more in-depth knowledge about DBNs and CHMMs.

## 4.2 Fusion

In human affective behavior analysis, modality fusion refers to combining and integrating all incoming unimodal events into a single representation of the observed behavior. Typically, multimodal data fusion is either done at the feature level in a maximum likelihood estimation manner or at the decision level when most of the joint statistical properties (maximum a posteriori) may have been lost [27].

Feature-level fusion assumes a strict time synchrony between the modalities. Therefore, it becomes more challenging as the number of features increases and when they are of very different natures (e.g., in terms of their temporal properties). Synchronization then becomes of utmost importance. Recent works have attempted synchronization between multiple cues to support feature-level fusion for the purposes of affect recognition, and reported greater overall accuracy when compared to decision-level fusion (e.g., [6], [20]).

In line with the aforementioned literature we employ feature level fusion as the cues and modalities employed in our studies are highly correlated. To this aim, features from all available cues and modalities are concatenated and fed into the static classifier. Moreover, to exploit the temporal correlation structure between the cues and modalities automatically via learning, we adopt model-level fusion based on Coupled Hidden Markov Models (CHMM).

# 5. EXPERIMENTS

In our experiments *feature representation* is chosen to be either frame-based or window-based. Frame-based representation refers to the features extracted for each audio (or video) frame as described in Section 3. In window-based representation the sequence is divided into 320ms long windows with 160ms overlap, and simple statistical features (mean and standard deviation) of the frame-level features are computed over each window. Note that window-based representation doubles the amount of features used, as for each feature the mean and standard deviations across a window are used for representation. Overall, the frame-based representation was chosen for its wide adoption within the research community, and the window-based representation was used due to its good performance reported in [15].

*Classification* is chosen to be either single cue or single modal, and multicue or multimodal, static or dynamic classification. The static classification scheme uses the features from one frame / window at a time and classifying it independently of the other frames. Taking the majority of the individual frame / window labels assigned by the classifier provides the label for the whole sequence. The dynamic classification scheme considers the sequential nature of the frames as in a time series. An entire sequence is fed to the classifier which outputs the label for this sequence.

## 5.1 Evaluation

Real world classification results can only be obtained if training cases and test cases are different. To achieve this and to make the most out of the data at hand, cross-validation is a common method for classifier evaluation. In our study, the cross-validation method is extended to subject independency. This is achieved by putting all cases belonging to the same subject into the same fold. Training is then performed by leaving out sequences contained in one fold, i.e., belong only to one subject . The left out subset is then used for testing. This procedure is repeated m times, where m is the number of subjects. each time leaving out data sequences from one different subject. In our study, for both cases we performed 10-fold subject-independent cross validation. This means that dor the first case study more than one subjects were left out for testing in each iteration. Classification accuracy, which is used as the performance measure in this study, is computed as the mean accuracy of the 10 repetitions.

## 5.2 Static Classification

Feedforward neural networks with one hidden layer are used as classifiers. The learning rate is set to 0.05 and the training is stopped when either the maximum number of epochs is reached (500 in our case) or the magnitude of the gradient is less than 0.04. The number of hidden neurons is defined by a 2-fold cross validation in the following way. The subjects in the 9 folds used for training are randomly divided into 2 groups. Then several networks are trained, with different numbers of hidden neurons, using only subjects from one group. They were tested on the other group. This was done for both groups and for all m folds. For each fold, the number of hidden neurons leading to the best performance is chosen for training a network on the entire training set and testing it for the subjects in the left-out fold. In each cross validation fold, all features used for training are z-normalized to a mean $\mu_{Norm} = 0$ and standard deviation $\sigma_{Norm} = 1$.

For Case Study 1, we wanted to investigate how frame-based representation affects static classifier's performance. To this aim we analyzed how single cues, namely facial expressions, head and shoulder movement, contribute separately to the aim of distinguishing between posed and spontaneous smiles using a static classification scheme. We then explored how fusing each pair of FaHe, FaSh, and HeSh features at the feature level affects the automatic discrimination of posed vs. spontaneous smiles. Finally, we fused all face, head and shoulder features at the feature level. Secondly, we wanted to see whether using a window-based representation affects static classifier's performance. We repeated the aforementioned experiments using the window-based representation for each cue. All results are presented in Table 2.

For Case study 2, we analyzed how single cues, namely Sp, PE and Fa, contribute separately to the aim of discriminating laughter from speech. We then explored how fusing each pair of features (FaSp, FaPE, and SpPE) at the feature level affects the classification. Finally, all features were fused together. The aforementioned experiments were then repeated using the window-based representation for each cue and modality. The main difference with Case Study 1 is that the audio and visual features are extracted at different frame rates. For frame-based representation synchronisation is achieved by upsampling the visual features (25 fps) by simply copying each visual feature so as to match the audio feature rate (50 fps). For window-based representation there is no need for synchronization since mean and standard deviation are computed for both modalities at the same rate, since the windows have the same length. All results are presented in Table 2.

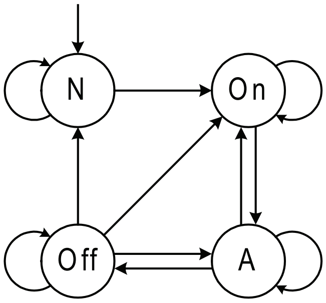

**Figure 3: Illustration of the HMM topology used for the face stream.**

## 5.3 Dynamic Classification

The temporal factors of a facial movement are described by four phases: neutral (there are no signs of muscular activation), onset (the muscular contraction begins and increases in intensity), apex (a plateau where the intensity reaches a stable level), and offset (the relaxation of the muscular action) [4]. Due to this nature of the face data we expect some of the states to only move forward and not come back. Therefore, the model size and state transition matrix for the face stream consists of four states (see Fig. 3), one for each temporal phase of neutral, onset, apex, and offset. This model has beed used for both Case Study 1 and 2. For Case Study 1, for head and shoulders, an ergodic HMM model (where all states are connected to each other) with two states is used, as the head and shoulders motion are modeled as either moving or non-moving. The implementation of the dynamic classification schemes has been done by using the Bayes Net Toolbox for Matlab [11].

The HMM model used for Case Study 1 can be described as follows:

- Number of states: 4 (neutral, onset, apex, offset) for face and 2 (active and inactive) for head and shoulders.
- Initial state probabilities: randomly generated.
- Initial state transition probability matrices: initialized randomly from a uniform distribution. The restricted transitions are then set to 0, and values are normalized.
- Density: continuous Gaussian distribution.
- Number of Gaussians per state: 1, 5.
- Weight for each of the Gaussian component: randomly generated from a uniform distribution.
- Covariance type: diagonal, value of each diagonal element is set to 100, and the rest are set to 0.

For Case Study 1, a similar structure is used for each of the cues that constitute the 2-chain and 3-chain CHMMs while adding transitions across cues. This is illustrated in Fig. 4.

Similar to static classification experiments, we first wanted to investigate how frame-based representation affects the classifier's performance. In the training stage, two HMM

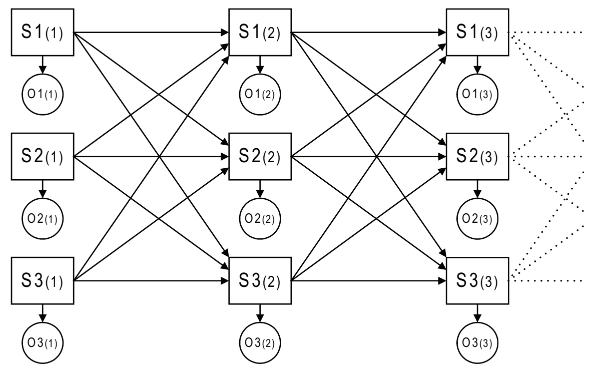

**Figure 4: Illustration of the 3-chain CHMM structure, (illustrated as a DBN) for discriminating between posed and spontaneous smiles, unrolled for 3 time-slices. Rectangle denotes a discrete node, circle denotes a continuous node and arrow denotes an intra-slice or an inter-slice connection. $S_1$ represents the face, $S_2$ represents the head, and $S_3$ represents the shoulder features.**

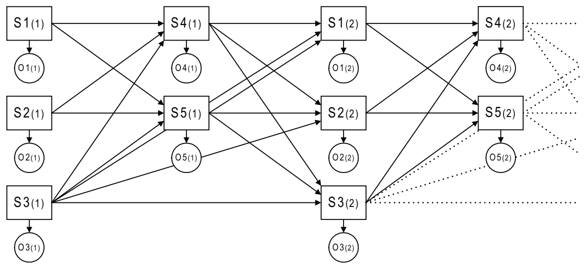

**Figure 5: Illustration of the 3-chain CHMM structure used for audio-visual laughter vs. speech discrimination unrolled for 3 time-slices. Rectangle denotes a discrete node, circle denotes a continuous node and arrow denotes an intra-slice or an inter-slice connection. $S_1$ represents the face, $S_2$ represents the spectral, and $S_3$ represents the prosodic features.**

/ CHMM classifiers were trained as the recognition task is to distinguish between posed and spontaneous smiles. Each classifier was trained with data belonging to its own class. During testing, a test sequence was fed to each of the classifiers separately, and one likelihood value was obtained from each classifier. The class membership of the test sequence was then decided according the classifier that provides the maximum likelihood.

We then temporally correlated each pair of FaHe, FaSh, and HeSh features by using a 2-chain CHMM classifier. Using the structure illustrated in Fig. 4, we trained a CHMM for each coupled cue : FaHe, FaSh, HeSh. We obtained one CHMM-based classifier for each set of FaHe, FaSh, and HeSh features, ending up with 3 separately trained models for each class of acted and spontaneous smiles. Finally, to correlate all face, head and shoulder features in time, we employed a 3-chain CHMM classifier.

We also wanted to see whether using a window-based representation affects a dynamic classifier's performance. We repeated the aforementioned experiments using the window-based representation for each cue. The results are presented in Table 2.

In order to obtain an insight on how different parameters affect the classification results, we extended our experiments by changing various parameters: number of states for different cues, ergodic vs. left-to-right transitions, 1 vs. 5 Gaussians. However, the overall recognition results did not change noticeably. Overall, our experiments suggest that the chosen topology with 1) the number of states (4 for face, 2 for head and shoulders), 2) connectivity of the states (as shown in Fig. 3), and 3) continuous probability densities with a low number of mixtures provide good results.

For Case Study 2, a 3 state ergodic model was used for the audio cues. For each of the cue and modality that constitute the 2-chain and 3-chain CHMMs the structure illustrated in Fig. 5 has been used. As can be seen from the figure, this structure differs from that used in Case Study 1 in terms of connectivity. For frame-based representation, the sampling rate of audio features is 50 fps, while it is 25 fps for video. In order to take into account this asynchrony in the model, an extra node is used in the audio stream as shown in Fig. 5.

In line with our previous experiments, we first wanted to investigate how single cues, namely prosodic features, spectral features and visual features, represented on a frame-basis, contribute separately to the aim of distinguishing between laughter and speech using a dynamic classification scheme. We then wanted to correlate each pair of FaSp, FaPE, SpPE features in time by using a 2-chain CHMM classifier. Finally, we wanted to correlate all prosodic, spectral and visual features in time by using a 3-chain CHMM classifier structure illustrated in Fig. 5. Again, for each class of laughter and speech we trained a separate CHMM model. We repeated the aforementioned experiments using the window-based representation for each cue and modality. Results are presented in Table 2 and analysis of all results is provided in section 6.

# 6. ANALYSIS AND CONCLUSION

Overall, from the experiments conducted in this study, and by looking at Table 2, it is possible to conclude that for machine analysis of human nonverbal behavior, both for static and dynamic classification, fusing data coming from multiple cues and modalities proves useful to the overall task of recognition.

When we analyze the experimental results for the task of automatic discrimination between posed and spontaneous smiles from multicue visual data we are able to state the following.

- When single-cue HMM classification is compared to that of paired 2- or 3-chain CHMM classification, accuracy increases significantly using the latter approach. This implies that coupling and correlating multiple cues in time enhances the discriminative power of dynamic classifiers.
- Face and shoulder cues seem to carry more complementary information about the meaning of the nonverbal message, compared to that of face and head, or head and shoulder cues. Therefore, combining information coming from face and shoulder cues improves recognition accuracy.
- Static and dynamic classification using the frame-based

**Table 2: The classification accuracy of static and dynamic classifiers for distinguishing between posed and spontaneous smiles (Case study 1) and discriminating between speech and laughter (Case study 2) using frame-based representation vs. window-based representation, and using single cue vs. multiple cue features. Results are averaged across subject-independent 10-fold cross-validation.**

| cues | NN | HMM/CHMM | NN | HMM/CHMM |
|---|---|---|---|---|
| | Frame | Frame | Window | Window |
| | | Case Study 1 | | |
| Fa | 0.75 | 0.53 | 0.80 | 0.68 |
| He | 0.71 | 0.51 | 0.76 | 0.51 |
| Sh | 0.72 | 0.52 | 0.74 | 0.50 |
| FaHe | 0.76 | 0.71 | 0.81 | 0.63 |
| FaSh | 0.80 | 0.78 | 0.83 | 0.69 |
| HeSh | 0.71 | 0.51 | 0.75 | 0.50 |
| FaHeSh | 0.79 | 0.78 | 0.82 | 0.66 |
| | | Case Study 2 | | |
| Fa | 0.84 | 0.84 | 0.87 | 0.84 |
| Sp | 0.92 | 0.58 | 0.93 | 0.85 |
| PE | 0.67 | 0.57 | 0.77 | 0.73 |
| FaSp | 0.94 | 0.82 | 0.97 | 0.94 |
| FaPE | 0.85 | 0.82 | 0.90 | 0.86 |
| SpPE | 0.93 | 0.83 | 0.93 | 0.91 |
| FaSpPE | 0.94 | 0.96 | 0.98 | 0.96 |

feature representation provide comparable results (80% vs. 78% for face and shoulder cues, 79% vs. 78% for all three cues).

- When combining all visual cues, static classification using frame-based feature representation provide similar results to static classification using window-based feature representation (79% vs. 82% for all three cues).
- When using window-based feature representation, static classification outperforms dynamic classification (83% vs. 69% for face and shoulder cues, 82% vs. 66% for all three cues).

Our experimental results show that in the task of discriminating between posed and spontaneous smiles static and dynamic classification using the frame-based feature representation provide comparable results. Additional experiments should be conducted to find whether the difference is statistically significant or not. Static classification using window-based feature representation outperforms dynamic classification. This is possibly due to the fact that window-based feature representation doubles the amount of features used (for each feature, the mean and standard deviations across a window are used for representation). Requiring much more traininig data than was available in this study for adequate training of dynamic classification schemes. While increasing the dimensionality does not seem to affect static classification, it visibly impedes the dynamic classification. More specifically, recognition accuracy goes from 78% for frame-based feature representation down to 66% for window-based feature representation. In general, it is known that dynamic classifiers are harder to train due to their complexity and number of parameters they need to learn [3]. As already said, they require more training samples compared to static classifiers.

An interesting result shown in Table 2 is that using all three visual cues (face, head and shoulders) did not provide the best recognition accuracy for discriminating between posed and spontaneous smiles. A possible explanation is that fusing features coming from all visual cues increases the dimensionality of the classification problem. Having fewer training samples than features per sample for learning the target classification may have led to under sampling or a singularity problem. To investigate this, in our future work we will apply dimensionality reduction or feature selection techniques.

For audio-visual discrimination between laughter and speech, the following conclusions can be drawn.

- The use of 2- or 3-chain CHMMs results in better performance than that achieved by single-cue HMMs.
- Both NNs and CHMMs perform similarly no matter which feature representation is used. It seems that the explicit temporal modeling provided by HMMs does not seem to be beneficial, since it achieves the same performance as a static model, i.e., NNs.
- The main difference between frame-based and window-based approaches, when used with NNs, is the performance of prosodic features. However, this is not surprising since usually pitch and energy information is much better encoded in a window than in a single frame.
- For HMM-based classification, both audio streams perform much better using the window-based feature representation than the frame-based approach. For face, both feature representations lead to comparable results.
- Using all audio and visual cues leads to the best performance for both types of feature representation and classification.

Overall, both static and dynamic classification schemes appear to provide very good results for automatic laughter-vs-speech discrimination. This in turn implies that the feature sets chosen are well able to represent the problem at hand making the task of audio-visual speech vs. laughter discrimination independent of the classifier choice. However, additional experiments should be conducted to find whether this conclusion is statistically significant or not.

Are dynamic classifiers better or worse than static classifiers for human nonverbal behavior analysis from multiple cues or modalities? As the research field has shifted its focus from static images to multimodal sequences, and from acted data to spontaneous and naturalistic data, more recent works in the field have considered that advanced data fusion methods relying on dynamic classifiers (e.g., tripled HMM [21], multi-stream fused HMM [28]) are better suited to the task of automatic human affective behavior analysis from multiple cues or modalities than static classifiers. However, our experimental results obtained for Case Study 1 and Case Study 2 show that the answer to the question we posed previously is not straightforward and depends on the feature representation (frame-based vs. window-based feature representation) and the task at hand. However, these findings have to be verified with more extensive experiments.

## 7. ACKNOWLEDGEMENTS

The work of Stavros Petridis and Maja Pantic leading to the presented results is funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Hatice Gunes is funded by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE).

## 8. REFERENCES

[1] J. A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal-Acoustical Society of America*, 110(1):1581–1597, 2001.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.01) (www.praat.org)). Technical report, 2005.

[3] I. Cohen and et al. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.

[4] P. Ekman. About brows: Emotional and conversational signals. In *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pages 169–248. 1979.

[5] D. Gonzalez-Jimenez and J. L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *IEEE Trans. Information Forensics and Security*, 2(3):413–429, 2007.

[6] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Tran. on Systems, Man, and Cybernetics-Part B*, 39(1):64–84, 2009.

[7] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In J. Or, editor, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218. Vienna, Austria, 2008.

[8] L. Kennedy and D. Ellis. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop*, 2004.

[9] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos. The ami meeting corpus.

[10] K. Murphy. *Dynamic Bayesian networks: representation, inference and learning.* PhD thesis, 2002.

[11] K. P. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33, 2001.

[12] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. *Proc. IEEE ICME*, pages 317–321, 2005.

[13] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Int'l Conf.Automatic Face and Gesture Rcognition*, pages 97–104, 2004.

[14] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *IEEE ICASSP*, pages 5117–5120, 2008.

[15] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In *Proc. ACM ICMI*, pages 37–44, 2008.

[16] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. In *Proc. ACM CIVR*, pages 329–337, 2008.

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

[18] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. of the IEEE*, 91(9):1306–1326, 2003.

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

[20] C. Shan, S. Gong, and P. W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *Proc. BMVC*, 2007.

[21] M. Song, J. Bu, C. Chen, and N. Li. Audio–visual based emotion recognition– a new approach. In *Proc. IEEE CVPR*, volume 2, pages 1020–1025, 2004.

[22] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proc. ACM ICMI*, pages 38–45, 2007.

[23] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In *Proc. ACII*, pages 139–147, 2007.

[24] T. Vogt and E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proc. IEEE ICME*, pages 474–477, 2005.

[25] T. Vogt, E. Andre, and J. Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *LNCS 4868*, pages 75–91, 2008.

[26] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int. J. Imaging Systems and Technology*, 13(1):85–94, 2003.

[27] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Tran. PAMI*, 31:39–58, 2009.

[28] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang. Audio–visual affective expression recognition through multistream fused HMM. *IEEE Tran. on Multimedia*, 10(4):570–577, 2008.