Computational Modelling to Track Human Emotion Trajectories Through Time

Ayesha Hakim, Stephen Marsland, and Hans W. Guesgen

Abstract—Having computers understand emotions has benefits for human-computer interaction, psychology, and behavioural analysis. This paper is an attempt to build computational models to track human emotions through time and analyse the trajectories followed by these emotions in a 3-D space. The paper is divided into three sections: first, building shape models to recognise how much of each emotion is expressed in a given frame; second, representing the frame in a 3-D emotion space; third, tracking the frames through time to analyse their paths and relationship with other emotions in the space.

Index Terms—Emotion recognition, Principal Component Analysis, Mahalanobis Distance, activation-evaluation space, von Mises.

1 Introduction

HUMANS tailor their interpersonal relationships by recognising emotions. This helps them cope with specific situations, such as realising when somebody else is annoyed. Research findings signify the importance of emotions in learning, decision making, and rational thinking [1]. Based on this, the ability to recognise and express emotions are essential for natural communications between humans. Nowadays, most of us spend more time interacting with computers than with other humans [1]. Computers control a significant part of our lives and we expect them to be reliable, predictable, and intelligent with rational judgment. We want them to be able to understand what we feel and to adapt accordingly. In order to communicate intelligently with us, computers will need the ability to recognise, express, and respond to our emotions.

Most of the research related to automatic emotion recognition focuses on giving computers the ability to recognise discrete basic emotions using static facial images of facial expressions. Some of which are discussed in Section 2. However, research in psychology and related fields has shown that our real-life emotions are not pure examples of one basic emotion, but a mixture of them, known as complex emotions. Basic emotions are combined with different 'weights' to give rise to several *blends* of emotions [1], [2], [3]. This paper starts with an introduction to a robust classification approach to map a set of facial points to the discrete basic emotion categories followed by a technique to map them to an appropriate emotion space and ended with the analysis of the trajectories followed by them while a person experiences a set of emotional states through time.

The paper is organised as follows: Section 2 presents an overview of the existing work related to the direct classification approach for basic emotion recognition and attempts to study emotion dynamics. The dataset chosen for training and testing the proposed model is explained in Section 3.1, along with some necessary data preprocessing steps. Section 3.2 presents a detailed analysis of dependent and independent shape models explaining the effects of speaker normalisation on overall emotion recognition accuracy. Section 3.3.1 describes an approach to map facial points into a 3-D emotion space based on a circular normal distribution. It compares the mapping of facial points classified by both dependent and independent shape models separately. Section 5 analyse the paths obtained by mapping the movement of facial points during continuous dialogs through time. The section also discusses our approach of defining the 'neutral face'. The paper ends with the description of results in Section 4 followed by discussion in Section 6.

2 Related Work

2.1 Basic Emotion Recognition

The classification of emotions using facial expressions in terms of categories has been the most common approach in the field of automatic emotion recognition. Since the 1970s, various techniques have been proposed to categorise facial expressions into a set of basic emotions or a set of facial action units (the smallest visually distinguishable movements on the face [4]), using either static facial images or a sequence of facial images.

- *Ayesha Hakim is with the Centre for Business, Information Technology & Enterprise, Wintec, New Zealand.*
  E-mail: ayesha.hakim@bzu.edu.pk
- *Stephen Marsland, and Hans W. Guesgen are with the School of Engineering and Advanced Technology, Massey University, New Zealand.*
  E-mail: s.r.marsland, h.w.guesgen@massey.ac.nz

2.1.1 Template-based techniques

In this technique, a template is defined for each emotion category and the unknown facial expression is compared to all the defined templates. The category of the unknown facial expression is decided on the basis of the best matched template.

To classify unknown facial expressions to six basic emotions, Huang and Huang [5] applied the statistical shape model also known as the point distribution model (PDM) [6] to extract the facial features. They calculated 10 Action Parameters (APs) to describe the position variations of certain points on the facial features. PDMs are the models used for the analysis of shape variation marked with landmark points [7]. They applied Principal Component Analysis (PCA) [8] to the APs of all training expressions and found that 90% of AP variations are covered by the first 2 eigenvectors. PCA is a commonly-used dimensionality reduction technique to analyse sets of data points in high dimensional spaces. This approach gives satisfactory classification accuracy, but the estimation of APs on the basis of gradient-descent-based shape parameters is a computationally expensive process. They reported an average computation time of seven minutes to analyse an image sequence approximately one second long. Also, the descriptions of the emotional expressions, defined in terms of facial actions, are incomplete.

Hong et al. [9] classified facial expressions into one of the six basic emotions or the expressionless (neutral) face using personalised galleries. It was assumed that two persons who look alike express emotions in a similar way. First, a labelled graph is fitted to the unknown face, then the best-matching person among those having the personalised templates is found by applying elastic graph matching [10]. The unknown facial expression is classified by using the personalised templates of the best-matched person. The system achieved 89% accuracy for the familiar faces (whose galleries were available), and 78% for unfamiliar faces. However, it only deals with full upright-frontal facial images and fails for profile or partially occluded facial images. The error rate of the system increases in the case of not finding the best-matched gallery, or the unavailability of the particular expression in the matched gallery.

Bartlett et al. [11] classified video frames into six basic emotions and neutral state by using Gabor energy filters [12], along with the recognition engines: AdaBoost [8], Support Vector Machines (SVM) [13], Linear Discriminant Analysis (LDA) [8], and feature selection techniques. Like [9], this system only deals with the frontal-view face images and does not account for the temporal dynamics (onset, apex, offset) of the facial expressions, or AUs.

In the work of Martins et al. [14], they used the Active Appearance Model (AAM) [15] to extract a facial geometry and Laplacian Eigen-Maps (LE) [16] to derive low-dimensional manifolds of that facial geometry. Martins et al. derived two types of manifolds: one which deals with the identity recognition, and the other for the person-specific facial expression recognition (expressions of six basic emotions and the neutral state). A multi-dimensional representation of a face can be represented by a single point in a multi-dimensional face-space and the variability of facial expressions can be represented as low-dimensional manifolds in that space [17]. The low-dimensional representation of facial changes in the face-space is a suitable approach to cover all possible variations of an emotional expression. However, the technique presented in this paper can only deal with the recognition of person-specific facial expressions of the familiar persons already present in the database and would fail for a face with unknown identity. Also, the 2D shape model ignores the detailed facial deformation which might improve the recognition of the minor variations of facial expressions.

Bansal et al. [18] used Latent Dirichlet Allocation [19] along with the Hidden Markov Model to classify facial image sequences to six basic emotions. Generally, the technique of Latent Dirichlet Allocation is used in natural text processing. Using this technique, each image sequence is assumed to be a document and each frame in the sequence is a word of that document. In this way, a set of image sequences is represented as the set of topics assigned to each frame. A Hidden Markov Model for each emotion was used to learn the sequence information of topics in image sequences, which is then used to classify image sequences to six basic emotion categories. However, the probability inference in LDA is a computationally complex process. When the number of emotion categories in a set of image sequence is small, the probability inference takes polynomial time, while in the case of several emotion transitions it is an NP-hard problem [20].

In [21], Costantini et al. demonstrated the role of upper and lower parts of the face in recognising emotions based on experiments including 74 participants. They compared the emotion recognition rate using the whole face, eyes, and mouth separately, and found that the eyes alone generate similar recognition rate to using the whole face, and higher recognition rate than using the mouth only. Several years ago, Bassili et al. [22] found that the upper part of face is associated with high recognition rate of negative emotions (e.g., sadness and fear), while the lower part of face yields high recognition accuracy of positive emotions (e.g., happiness). These results motivate the development of computational models of the full, upper and lower parts of the face separately, however in the reviewed literature we could not find any systematic experiments modelling and analysing the upper, lower, and full face templates separately for the task of emotion recognition.

2.1.2 FACS-based techniques

A very common method for measuring facial expressions in behavioural science is the Facial Action Coding System (FACS) [4], [23]. FACS is a scoring system defined for expert human observers, not computers. It aims to provide objective measures of facial activity to assist behavioural science analysis of the face. Ekman and Friesen defined 46 Action Units (AUs) that describe the smallest visually perceptible facial movements. They determined the effect of contraction of each of the facial muscles on the visible appearance of face by using knowledge of facial anatomy. The system can be used to describe any facial movement (observed in images or videos) in terms of anatomically based action units. This system has been used, for instance, to demonstrate differences between genuine and simulated pain [24], differences between the facial signals of suicidal and non-suicidally depressed patients [25], and differences between when people are telling the truth versus lying [26].

A lot of researchers have tried to automate FACS by recognising facial action units in images and/or videos, e.g., Donato et al. [27] used Gabor wavelets and Independent Component Analysis (ICA) to recognise eight individual AUs and four combinations of AUs. Cohn et al. [28] used facial feature point tracking and discriminant function analysis to recognise eight individual AUs and seven combinations of AUs. Sixteen AUs were recognised by Tien et al. [29] using lip tracking, template matching, and neural networks in nearly-frontal facial images. Braathen et al. [30] recognised three AUs using particle filtering, Gabor wavelets, SVM, and HMM in facial images with varying head poses. Mahoor et al. [31] measured the intensity of AU12 and AU6 in videos captured from infant-mother interaction by using the SVM.

Most of the automatic systems using FACS are trained and tested on posed expressions where actors are asked to voluntarily contract specific muscles corresponding to certain AUs. Also, reliable FACS-coding of facial images/frames is a long tiring process which needs FACS-certified coders to spend hours to code just a few seconds of video. Although FACS is a promising approach, in reality it is not always possible to locate the action units in each image/frame due to changes in lighting conditions, occlusions caused by the facial hairs and glasses, and poor image quality. These problems also exist for other feature extraction techniques using facial images or videos.

2.1.3 Rule-based techniques

The rule-based techniques classify the unknown facial expressions to emotion categories based on rules applied to the movement of facial action units or the facial definition parameters (FDP) which define the shape of the face [32].

Pantic et al. [33] defined rules based on FACS to classify facial expressions to 31 action units as well as six basic emotions. The rules were applied to the model deformation parameters calculated by taking the difference between the detected model features and the same features detected in the neutral face of the same person. The system reported high average accuracy, but is limited to just static images of posed expressions.

Zhou et al. [34] divided the facial expressions into three categories based on the deformation of mouth region. Anger, sadness and disgust come in the first category, happiness and fear in the second category, and surprise in the third category. Classification to the six basic emotions was done by applying rules to the displacement of the key points of eyes, eyebrows, and mouth extracted by using an Active Appearance Model. As in [33], this system has also been tested on static facial images of posed expressions without considering the temporal dynamics of facial expressions.

One of the limitations of rule-based systems is that the rules are only applied properly if the feature set is extracted correctly from images. The reliable extraction of facial features is a difficult task for images which are captured under complex backgrounds, varying lighting conditions, showing spontaneous expressions, uncontrolled head movements, and/or are partially or fully occluded.

2.2 Complex Emotion Recognition

After decades of research in emotions, researchers have shown that in everyday interactions people exhibit non-basic, intermediate, or complex emotional states which are related to each other in a systematic manner. Based on this, it is not appropriate to assign a single independent categorical label to complex emotions which are mixtures of more than one emotion category. Therefore, there is a shift in research towards emotion space representations, where emotions are mapped into an emotion space either in quantised levels or along a continuum, in part to recognise the fact that emotions are a continuous phenomenon and in part to enable complex emotions to be identified without requiring labels [**?**].

2.2.1 Quantised approaches

In this approach, the emotion attributes (e.g., valence and activation) are quantised into an arbitrary number of levels or intensities. In this approach the most common way is to reduce the emotion classification problem to a two-class problem (positive versus negative and active versus passive) or to a four-class problem (quadrants of 2D activation-evaluation space).

Ioannou et al. [35] classified emotions in video frames to the neutral state, the six basic emotions as well as three quadrants of the activation-evaluation space (there was no emotion lying in the fourth quadrant). They defined fuzzy rules based on the variations of Facial Action Parameters (FAP). FAPs were defined by Pandzic et al. [32], and are closely related to muscle actions, and represent a complete set of basic facial actions along with head motion as well as eye, tongue, and mouth control. The system is dependent on the robust extraction of FAP variations, which is quite difficult in the case of naturalistic data. Also, the classification to the quadrants of the activation-evaluation space does not give much information about the emotional state, since each of them contains emotions expressed with highly varying features (e.g., anger and fear both lie in the same negative/active quadrant) [**?**].

Shin et al. [**?**] used facial expressions as a clue from the Korean facial expression database [**?**] selecting 252 static images for training and 66 images for testing. The images consists of 11 expressions of 6 subjects (3 males and 3 females). They classified expressions into 9-point scale for valence and activation using manifold learning for the feature extraction of facial expressions. The system reported 90.9% accuracy for valence and 56.1% for activation. However, the evaluation has been done on a very small testing set, which raises doubts about the reliability of the reported performance accuracy.

Following the quantised approach of emotion classification, most of the published work in the literature uses multiple clues, e.g., Karpouzis et al. [**?**] used clues from facial and hand gestures as well as vocal information to classify naturalistic emotions into neutral state and 3 quadrants of the activation-evaluation space. Similarly, Caridakis et al. [**?**] discriminated emotions into 5 classes (neutral state and four quadrants of activation-evaluation space) using a feed-forward back-propagation network. The system performed decision-level fusion of two visual modalities, i.e., facial expressions as well as hand and body gestures using the Sensitive Artificial Listener (SAL) dataset [**?**]. As in [35], these two approaches also classify emotions to the quadrants of the activation-evaluation space, which makes it difficult to differentiate between emotions within the same quadrant.

Kulic et al. [**?**] used facial muscle contraction, heart beat, and perspiration rate to classify emotions into low, medium, and high levels of valence and activation. They implemented a Hidden Markov Model to estimate emotion attributes of 36 human subjects during human-robot interaction. The system is based on the physiological signals gathered by using different sensors on the body, which makes it invasive and unsuitable for capturing naturalistic emotion data.

Some efforts have been made on emotion classification based on quantised approaches using motion capture signals. For example, Wöllmer et al. [**?**] used the facial markers information as well as audio clues from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [36] to classify a set of utterances to three levels (negative, neutral, and positive) in terms of two emotion attributes: valence and activation. They used Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Networks (RNN) which consists of two neural networks: the first processes the sequences forwards and the other processes it backwards. In this way, the BLSTM RNNs have access to past and future data points in the sequence. However, their system is based on the segmented utterances (not continuous conversations), which often contains long breaks due to several silent frames in a sequence. The performance may be improved by utilising this missing data while considering the contextual information.

The quantised approach to emotion recognition is a step forward to the direct classification approaches in a sense that it gives more information about the emotion attributes and intensity. Also, it enables the classification of relatively large numbers of discrete emotions as compared to just the basic emotions or a few action units. On the other hand, it is a simplified problem in terms of emotion space approaches that can represent a set of discrete emotions in terms of emotion attributes. Although it gives an overall idea about the nature of emotional states by placing them into quadrants or assigning some level of intensity, it still leaves them indistinguishable in terms of fine categories. Nevertheless, neither of these two approaches is able to analyse the dynamics of a large number of natural emotions and their relationships to each other. Moving from one basic emotion to another in the case of categorical description and from one quadrant to another in the case of quantised labelling would not make much sense in real life scenarios. In the following section, we describe the continuous approaches for emotion recognition by mapping them into emotion spaces.

2.2.2 Continuous approaches

In real life, we come across continuous complex emotions rather than discrete basic categories or quantised labellings. The term 'continuous' refers to the uninterrupted sequence of emotions, which are dynamic in terms of the changing facial patterns, the speed of onset, apex and offset movements, their intensity and duration. There are fuzzy boundaries between emotional states which are too vague to be separated. Based on this, research in computer vision is shifting towards continuous emotion recognition by mapping emotions into the emotion space continuum to explore the dynamics of complex emotional states. However, the research in this area is still in its infancy.

In 2008, Kanluan et al. [**?**] used facial expression and audio clues to classify emotions in terms of three continuous emotion attributes (valence, activation, and dominance). They used SVM Regression (SVR), *k*-Nearest Neighbor estimator and a fuzzy logic estimator to estimate emotion attributes using both modalities separately. The emotion estimation for each modality was fused at a decision-level using a weighted linear combination. The system used the VAM corpus [?] (videos recorded at 25 frames per second) in which the data annotation was done on a 5-point scale (mapped to a scale of [-1, +1]) for each emotion attribute. The results show that valence was best estimated using the visual clues, while the estimation of activation and dominance was best using the combination of visual and audio information. These results might be influenced by the choice of facial features used in the training, i.e., eyes and lips. The lower face (i.e., lips/mouth) region is not considered as a good estimator of emotions, especially in *talking* scenarios as in the chosen dataset [21].

Hupont et al. [?] in 2010 presented a system for mapping complex emotions to the activation-evaluation space based on the positions of discrete basic emotions and the neutral state in that space. The corresponding angles, valence, and activation values of basic emotions and the neutral state are listed by Whissell in [?]. The basic emotions were classified based on the distances and angles between a set of facial points. This was done by using a combination of classifiers implemented by the Waikato Environment for Knowledge Analysis (WEKA) tool [?]. The unknown emotion was mapped to the activation-evaluation space using a weighted linear combination of the position (in x, y coordinates) of the basic emotions in that space, although it is unclear from their paper how this was actually done. However, it is somehow based on the valence, activation, and angular values listed by Whissell. The weights are associated with the confidence value of classifying an unknown facial expression using a set of basic emotion classifiers. The authors reported that it is possible to measure the intensity of emotions based on their attributes (valence and activation) values, but do not say how.

In 2011, Nicolaou et al. [?] used the Output-Associative Relevance Vector Machine (OA-RVM) regression framework to predict valence and activation from naturalistic facial expressions. The presented framework is based on learning non-linear input and output dependencies in the emotion data. The system was tested on the SAL dataset reporting high prediction accuracy as compared to both RVM and SVM. The proposed framework is quite robust for continuous emotion classification to valence and activation, however further analysis of temporal dynamics is needed in order to understand the correlation between these two emotion attributes.

In the work of Dahmane et al. [?] in 2011, they used Gabor energy filters to extract facial features and multi-class SVM to classify emotion in terms of four attributes: valence, activation, expectancy, and power. The results show quite low recognition accuracy for valence (48%) and power (36%), reaching an overall accuracy of 51.6% for all emotion attributes using the SEMAINE dataset [?]. The classification was done by sampling video frames at an interval of 10 frames (videos were recorded at 50 frames per second) which might lose some information.

In 2012, Martinez et al. [?] presented a model (without actual implementation) consisting of C distinct continuous spaces one for each basic emotion category. The spaces are linearly combined to represent *blends* of emotions. The intensity of each contributing emotion in the blend defines the weight of each emotion in the combination. The idea of representing complex emotions by mixing the basic emotions is similar to our approach for complex emotions recognition, but the linear combination of emotion spaces proposed in this model is not suitable to characterise complex emotions. The reason why linear operations are inappropriate to analyse emotion spaces will be illustrated in Section 5.

Some efforts have emerged for continuous emotion recognition using audio modality, e.g., [?], [?], [?], head gestures, e.g., [?], and thermal signals, e.g., [?], [?]. However, to the extent of our knowledge the continuous emotion recognition and analysis has not been attempted yet using motion capture data.

2.3 Analysis of Temporal dynamics of emotions

Temporal dynamics of emotions play an important role in the proper interpretation and understanding of emotions [?]. The information about facial expressions through time helps to interpret the relationships between emotions, such as how the intensity of one emotion changes while transitioning from one state to another, and what paths the emotions follow during emotion transitions. In the literature, very few efforts have gone into detecting and analysing temporal dynamics of emotions, focusing only on detecting whether a certain facial expression or a combination of AUs is in its onset, apex, or offset phase.

In the work of Valstar et al. [?], they detected the presence of any of 15 AUs per frame along with some aspects of their temporal dynamics, i.e., whether the detected action unit is in its onset, apex, or offset phase and the total duration of activation of that AU. They based this analysis on the tracking data of 20 facial points detected using the GentleBoost [?] template. The action units and their temporal dynamics were classified using Support Vector Machine (SVM) [13] on the posed video sequences from the MMI dataset [?]. This system is helpful to study the movements of some of the facial muscles, but is of less help for facial expression classification. One of the reasons is that the number of AUs detected by this system is too small to be mapped to a wide range of facial expressions. Also, the system is evaluated on the posed neutral-onset-apex-offset-neutral sequences of AUs, that makes it unsuitable for practical applications.

Gunes et al. [?] detected the emotion segments by finding the start and end of the neutral-onset-apex-

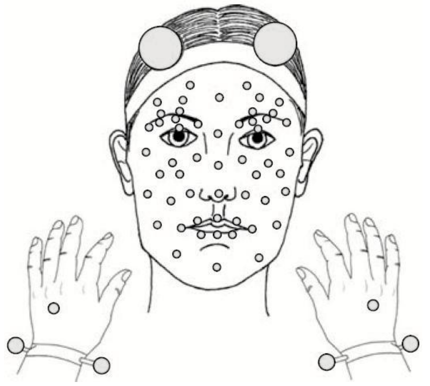


Fig. 1: Marker Layout used in recording for IEMOCAP dataset [36].

offset-neutral phase from face and body videos by comparing each frame to the reference (neutral) frame as well as consecutive frames. Using the apex frame out of the detected emotion segment, they detected six basic emotions as well as boredom, anxiety, uncertainty, puzzlement, and surprise (positive, neutral, and negative) using facial expressions and body gestures. The system was tested on the Bimodal Face and Body Gesture (FABO) [**?**] dataset, where subjects were asked to act out certain emotion-eliciting scenarios in laboratory settings. The system detects the apex from the posed neutral-onset-apex-offset-neutral sequence of emotions, which has limited its use in naturalistic scenarios.

To the best of our knowledge, so far none of the computational studies has focused on the analysis of temporal dynamics in order to understand the relationships between different emotions, the paths followed by emotions while moving from one state to another (i.e., emotion trajectories), or to study the change in emotion intensity through time.

3 Methodology

3.1 Dataset

We have selected a dataset (the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset from the Speech Analysis and Interpretation Laboratory (SAIL) at the University of South California (USC) [36]) where ten actors were recorded during both scripted and improvised sessions with a set of reflective markers on their face and hands. The layout of the 53 facial markers, 2 hand markers, 4 wristbands markers, and 2 headbands markers is shown in Fig. 1. The facial markers provide detailed motion capture information about their facial expressions. High speed cameras capturing 120 frames per second were used to record the actors and the 3D positions of the facial markers was tracked with very high accuracy.

The released version of the dataset contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. It consists of dyadic sessions (5 sessions with 2 subjects each), where actors perform improvised hypothetical scenarios and selected emotional scripts, specifically selected to elicit emotional expressions (happiness, sadness, anger, frustration, and neutral state).

During each interaction, only one of the two actors was motion captured at a time while performing improvised and scripted dialogs. Each dialog consists of almost 25 utterances/turns of each actor, where an utterance is a sentence or similar period during which one actor talks continuously. The dataset contains ten thousand and thirty nine utterances where each utterance consists of almost 11.4 words with an average duration of 4.5 seconds.

Within each dialog, each utterance of the two actors was annotated by three independent human evaluators into categorical labels (neutral, happiness, anger, sadness, surprise, disgust, fear, frustration, and excitement) as well as psychological data about emotion intensity (valence, activation, and dominance).

These two approaches cover the categorical as well as continuous attributes of emotional representation. The emotion content of the dataset was annotated by the evaluators into categorical labels using the ANVIL tool [37], after sequentially watching the videos. In this way, the evaluators had information from audio and video channels as well as previous utterances in the dialog for assessing the emotional content of each utterance. It was assumed that within an utterance, there is no emotion transition (e.g., from happy to excited), however, the evaluators were allowed to select more than one emotion category to describe mixtures/blends of emotions, which are more common in natural communication. The estimated confusion matrix between the assigned categorical labels shows that there is an overlap between happiness and excitement as well as anger and frustration. Neutral, disgust, and anger are often get confused with frustration. Also, sadness is often confused with frustration and neutral.

The evaluation-level limitation of the selected dataset is associated with its utterance-based annotation technique. Assuming the emotional content did not change much within an utterance (duration ≈ 4.5 seconds), the same label (or mixture of labels) was assigned to each frame in that utterance. The 'silent' frames and those containing sounds of active listening like 'mmhh' were not annotated at all. Another problem of the chosen dataset is that the human evaluators were sometimes inconsistent in labelling the data, since each one of them perceived and evaluated the emotions associated to an utterance in his own way. Due to the subjective nature of emotions, the inter-evaluator inconsistency is a common problem of all emotion datasets.

3.1.1 Data Preprocessing

We used the locations of the marker points in 3D as the basis of our analysis, and chose $5,000$ frames of each of five emotions (happiness, excitement, anger, frustration, and sadness) and neutral state for each of the ten actors to form a training set of $300,000$ frames. It should be noted that the marker points were already aligned to make the nose marker at the center of each frame that removed any translation effects. The rotational effects were compensated by multiplying each frame by a rotational matrix. For details about markers alignment, refer to [36].

Each utterance was labelled by the three expert human evaluators in terms of discrete categories as well as emotion attributes (valence and activation). For the training set, we took frames from the utterances where all three experts agreed. We used six emotions rather than the full nine as for the missing emotions (disgust, surprise, and fear) there was insufficient data, sometimes as little as $2,000$ frames in total. Out of the six selected emotions, two (frustration and excitement) are the candidate basic emotions [38], [39]. The selection of the testing set would be described in Section 3.2 depending on the type of model (dependent or independent) used for training. For the testing set, there was no such condition of agreement by all three experts while choosing the frames.

Each frame of the dataset contains the motion capture information of 61 markers in 3 dimensions, so the training data was of size $300,000 \times 183$ dimensions. We reduced the dimensionality of the data for each frame in three ways:

1) Markers not on the face (such as the head and hands) were excluded.
2) Markers that did not move significantly (such as eyelids and nose) were removed.
3) Sets of markers that moved together (such as, points on the chin and forehead) were replaced by a single point at the centre of the set.

As a result of these simplifications each emotion frame is represented by 28 markers points covering the forehead, eyebrows, eyes, cheeks, lips, and chin (Fig. 2). The location of marker points are in 3D, making it an 84D vector.

3.2 Shape Models

Based on the data from IEMOCAP we had sets of facial points with an associated emotion label. As described in Section 3.1.1, we are using the 3D locations of 28 marker points on the face for emotion recognition and analysis. We then develop two types of models: a model of only one actor at a time that we called the **speaker-dependent model**, and a model of nine out the ten actors that we called the **speaker-independent model**.

The **speaker-dependent model** is trained on 5000 frames of each emotion of one actor only to form a training set of $30,000$ frames. We then used principal component analysis (PCA) to develop a face model of the given training set based on the method described in [**?**]. We noticed that the first 7 PCs covered 93% of the total variation of the training data out of which the first PC, which covers almost 50% of the total variation, was describing the upward and downward movement of the mouth points. This movement of lips was experimentally shown to be highly correlated with talking [**?**], which is not directly connected with emotion recognition, and not much else, and so we discarded the first PC. Consequently, we chose to use six PCs (2-7) of the face model for our analysis. For details about the effects of each principal component on the mean face, refer to [**?**].

We transformed the training data into the 6-D space of the selected six principal components. Each datapoint was then labelled with the majority vote of the three human experts, so that the training set consists of 30,000 points, each labelled with one of six emotions in the 6-D space.

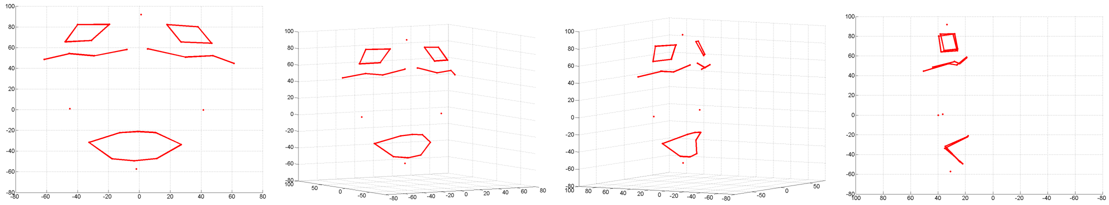
To evaluate the speaker-dependent face model, we chose the continuous testing frames of the same actor on which the model was trained. That was the ideal situation as there was only one face involved leading to no speaker variability and the first PC correlated to talking was removed leading to no lexical variability. For classification of a test frame, it was transformed into the 6-D space of the dependent face model. We then computed the Mahalanobis distance between the test frame and the six emotion clusters. In this way, we get a set of six distances; one for each emotion in the model space. We use these distances to map the emotion into the activation-evaluation space 3.3.

The **speaker-independent model** is trained on 5000 frames of each emotion of nine actors to form a training set of $270,000$ frames. We then again used PCA to develop a face model of the given training set and noticed that the first 15 PCs covered 90% of the total variation of the training data. Unlike speaker-dependent model, the first PC was not describing talking movements rather was related to the face variations of nine different speakers. However, the speaker variation was not too well defined in the first few PCs to discard.

To evaluate the speaker-independent face model, we chose the continuous testing frames of one of the ten actors that was not included in the training set. For classification of a test frame, it was transformed into the 11D space of the independent face model. We then computed the Mahalanobis distance between the test frame and the six emotion clusters to get a set of six distances; one for each emotion in the model space. We use these distances to map the emotion into the activation-evaluation space 3.3.

3.2.1 Speaker Normalisation

The speaker variability in the **speaker-independent model** is removed by reference-based speaker normalisation that is quite similar to the technique used in [**?**], [**?**].



(a) rotation: 0° (b) rotation: 30° (c) rotation: 60° (d) rotation: 90°

Fig. 2: The 28 marker points in 3D used for emotion recognition and analysis.

The neutral interaction of the female subject in the first session of the database is selected as the reference speaker. The technique is about matching the first and second order statistics of facial points movements across all the speakers with respect to the reference speaker. The markers of other speakers are mapped into the markers' space of the reference speaker. Equation 1 describes this speaker normalization technique. The i^{th} marker of the speaker s in the direction $d \in X, Y, Z, (m^{s}_{i,d})$, is transformed to match the reference speaker (ref), where μ and σ are the mean and standard deviation of the markers.

$$m^{s'}_{i,d} = (m^{s'}_{i,d} - \mu^{s'}_{i,d}) \times \frac{\sigma^{ref}_{i,d}}{\sigma^{s}_{i,d}} + \mu^{ref}_{i,d} \qquad (1)$$

After speaker normalisation, we retrain the **speaker-dependent model** on the normalised $30,000$ frames containing all emotions of one actor only and used PCA to develop a normalised face model of that actor. We repeated the same experiment on the normalised $270,000$ frames of all excluding one speaker to get a normalised **speaker-independent model**. Section 4 shows the comparison of all face models: speaker-dependent face model, speaker-independent face model, normalised speaker-dependent face model, and normalised speaker-independent face model.

3.2.2 Classification

For classification based on each face model separately, we replaced each cluster (i.e., set of points labelled as one emotion) with the mean of that set, and also computed the covariance matrix (spread) of the data. We therefore ended up with six datapoints representing the mean of each set and an associated covariance matrix.

For classification of a test frame, it was transformed into the transformed space of each model separately. We then computed the Mahalanobis distance between the test frame and each of the six emotion clusters. In this way, we got set of six distances; one for each emotion in each model space. We then labelled the test point with the label of the cluster that it is closest to. The Mahalanobis distance not only uses the mean, but also takes into account the spread of the data to compute a distance. It is formulated as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}$$

where $\mathbf{x}$ is the (4D) column vector of the test frame, μ is a column vector of the mean and Σ is the 4×4 covariance matrix for an emotion. If the covariance matrix is set to the identity matrix, then the Mahalanobis distance reduces to the Euclidean distance [8], [40].

Computing the Mahalanobis distance is a computationally expensive process which calculates the covariance matrix and then its inverse. For efficient computation, we have used Matlab's *mahal* from the Statistics toolbox to compute the Mahalanobis distance of each frame to each of the six emotion clusters.

3.3 Mapping emotions into the activation-evaluation space

There are two steps required to map the representation of the facial points of an image frame into the activation-evaluation space: represent the basic emotions as points within that space, and then position each frame (using the six distances to the basic emotions). The first of these steps uses the training data, which is assumed to represent each of the six basic emotions (all three experts agreed on their labels), while the second uses the testing data. We used the same technique to calculate the position of mean emotions and test frames on the emotion-space.

For each of the four face models, we calculated the position of the mean emotions separately based on the respective training data set. For each training set separately, we developed one shape model and calculated the distance of each frame from each of the five emotion clusters$(\beta_{ang}, \beta_{fru}, \beta_{hap}, \beta_{exc}, \beta_{sad})$ and the neutral state (β_{neu}).

Each frame is assumed to be a combination of the basic emotions, and so we needed to calculate the weighted average of basic emotions, where the weights correspond to the classification confidence of test frames for each basic emotion. We calculated the weight of each emotion and the neutral state as follows,

$$\gamma_j = \left(\frac{\beta_j}{\sum_{i=1}^{6} \beta_i} \right)^{-1} \qquad (2)$$

where j is the number of emotions and the neutral state.

To calculate the weighted average of basic emotions for each frame, we modelled the distribution of each basic emotion as a von Mises distribution and constructed a mixture model of them. This is described in Section 3.3.1.

To plot the mean of each emotion in the emotion-space, we calculated the x-coordinates and y-coordinates for each frame as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3}$$

where n is the number of frames of each emotion, x_i and y_i would be calculated using the mixture of von mises distribution as discussed in Section 3.3.1.

It should be something like this. or should mention mixture model here:

$$\begin{aligned} \bar{X} &= \sum_{i=1}^{j} \gamma_i \times cos(\theta_i) \\ \bar{Y} &= \sum_{i=1}^{j} \gamma_i \times sin(\theta_i) \end{aligned} \tag{4}$$

where j is the number of emotions and the neutral state.

We calculated the mean magnitude of the resultant vector for each of the six emotions separately as follows,

$$\bar{R}^2 = \bar{X}^2 + \bar{Y}^2 \tag{5}$$

The mean of $\bar{R}$ corresponds to the mean intensity of each emotion in the activation-evaluation space.

To calculate average angles it is not appropriate to use the linear statistical mean average due to the reasons discussed in Section 5. Therefore, we chose to calculate the geometric mean to get the angular position of each of the mean emotions in the activation-evaluation space. The computation of the mean direction is as follows:

$$\mu = \begin{cases} \tan^{-1}(\bar{A}/\bar{V}) & \bar{A} > 0, \bar{V} > 0 \\ \tan^{-1}(\bar{A}/\bar{V}) + \pi & \bar{V} < 0 \\ \tan^{-1}(\bar{A}/\bar{V}) + 2\pi & \bar{A} < 0, \bar{V} > 0 \end{cases} \tag{6}$$

μ is the mean direction and $\bar{R}$ corresponds to the mean intensity of emotions and neutral state in the activation-evaluation space. This gave us locations for the basic emotions and the neutral state. The location of each test frame is also calculated by the using equation 5 and 6.

3.3.1 von Mises Mixture Model

A circular variable θ is said to have a von Mises distribution if the probability density function is given by:

$$m(\theta; K, \mu) = \frac{1}{2\pi I_0(K)} e^{[K\cos(\theta-\mu)]} \tag{7}$$

where $0 \le \theta < 2\pi$, $K > 0$ and $0 \le \mu < 2\pi$.

The parameter μ is the mean direction and K is the concentration parameter, which is analogous to the (inverse) variance: the density at the mode depends on e^{2K} and the larger the value of K, the greater is the clustering around the mode. The distribution is uni-modal and symmetric about μ. $I_0(K)$ is a normaliser to turn this into a probability density function and consists of a modified Bessel function of the first kind of order zero [?]:

$$I_0(K) = \sum_{r=0}^{\infty} \frac{1}{r!^2}\left(\frac{1}{2}(K)^{2r}\right) \tag{8}$$

Although each emotion class is uni-modal, we cannot fit one von Mises distribution to the full data as it is the mixture of six different emotion classes. Such multi-modal distributions may be regarded as mixtures of uni-modal distributions [?]. We used a finite mixture model of six uni-modal von Mises distributions, given by:

$$M = \sum_{j=1}^{6} \gamma_j m_j(\theta) \tag{9}$$

where γ_j are non-negative weights that sum to one. We have already calculated the mean direction (μ_j) of each of the six reference emotions in the space using Eq. (6), and the methods of estimating K_j and γ_j are described in the following section.

3.3.2 Estimating the Parameters of the Mixture Model

There are several ways to estimate the parameters on which the mixture model depends [?]. We have used the usual maximum likelihood estimate for K_j. However, the weights γ_j of each emotion model are estimated by using the distances to the six emotions calculated by the shape models.

3.3.2.1 Estimating the concentration parameter: The concentration parameter K_j is estimated by using the Fisher equation [?]:

$$\hat{K}_{ML} = \begin{cases} 2\bar{R} + \bar{R}^3 + 5\bar{R}^5/6 & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + 0.43/(1-\bar{R}) & 0.53 \le \bar{R} < 0.85 \\ 1/(\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}) & \bar{R} \ge 0.85 \end{cases} \tag{10}$$

$\hat{K}_{ML}$ may be biased if the sample size (n) and $\bar{R}$ are small (specially when $\bar{R} < 0.45$). For this reason, if $n \le 15$, the following estimate is to be preferred:

$$\hat{K} = \begin{cases} \max(\hat{K}_{ML} - 2(n\hat{K}_{ML})^{-1}, 0) & \hat{K}_{ML} < 2 \\ (n-1)^3\hat{K}_{ML}/(n^3+n) & \hat{K}_{ML} \ge 2 \end{cases} \tag{11}$$

This is a standard approach for estimating the concentration parameter [?], [?].

3.3.2.2 Estimating the weights: We have already calculated the Mahalanobis distance of each test frame to each of the six basic emotions using the shape models. We want to position each test frame in activation-evaluation space using the positions of the basic emotions. However, the Mahalanobis distance is an unsigned quantity and so we do not know the direction between the test frame and the mean of each of the clusters of basic emotions. Since we have assumed that each emotion lies along a radial line in the activation-evaluation space we want to compute the intensity of each of the basic emotions as a component of the complex emotion. We did this starting at the position of the basic emotion and then by applying a simple rule to move along that radial line: if the distance of the test frame from neutral is less than the mean of a particular emotion, then the distance of the test frame from that emotion should be towards neutral i.e., its intensity decreases and comes close to neutral and vice-versa. We convert these distances to weights (γ_j) using equation 2.

4 Results

Table 4 lists the mean locations of the five emotions and neutral state in the activation-evaluation space based on different training sets. For the training data of the first speaker only, the mean intensity of five emotions and neutral state remain same with and without normalisation. For the training data of all speakers together, the mean intensity of emotions decreases without normalisation caused by the speaker variation leading to confused emotion recognition. The mean intensity increases after reference-based speaker normalisation.

We observed that the estimated directions of each emotion based on training data are quite close to those specified by Whissell in [?], except that of neutral state. According to the values listed by Whissell, neutral lies close to the centre of circle that corresponds to both valence and activation close to zero. In our data, since the direction of neutral is calculated as weighted average of all emotions its more inclined towards anger/frustration. One reason is that there are more emotions on the side of negative emotions than the positive emotions. Other reason is based on the fact that mostly the human evaluators misinterpreted neutral as frustration or sadness. This direction of neutral state does not effect the position of other emotions because the neutral state represents the 'no-emotion' state whose position should be uncorrelated with that of all emotional states.

Based on these positions for the basic emotions we were now able to compute the parameters of the mixture model and test it using initially single utterances with only one labelled emotion, and then full dialogs with several emotion transitions. We use the valence and activation values assigned to each utterance by three human experts to estimate the ground truth direction and intensity of emotion associated with that utterance, which we can compare to our results [?].

In order to test the goodness of fit of two sample distributions (ground truth and model estimation), we also used Kuiper's test, which is a circular analogue of the Kolmogorov-Smirnov test [?], [?]. Let $F(\theta)$ denote the continuous cumulative distribution function (cdf) of each of the emotions in the mixture model separately and $S_n(\theta)$ be the ground truth cdf (referred to as the empirical cdf). The Kuiper's statistic is defined as,

$$V_n = D_n^+ + D_n^- \tag{12}$$

where

$$D_n^+ = \max[S_n(\theta) - F(\theta)], \quad D_n^- = \max[F(\theta) - S_n(\theta)] \tag{13}$$

D_n^+ and D_n^- are the discrepancy statistics, where D_n^+ is the maximum vertical distance of $S_n(\theta)$ from $F(\theta)$ when the distance is measured above $F(\theta)$, while D_n^- is the maximum distance measured below $F(\theta)$. Both statistics D_n^+ and D_n^- depend on the choice of zero direction, but their sum (V_n) is invariant under rotation. This makes the Kuiper's statistic equally sensitive at the median as well as at the tails [?], [?].

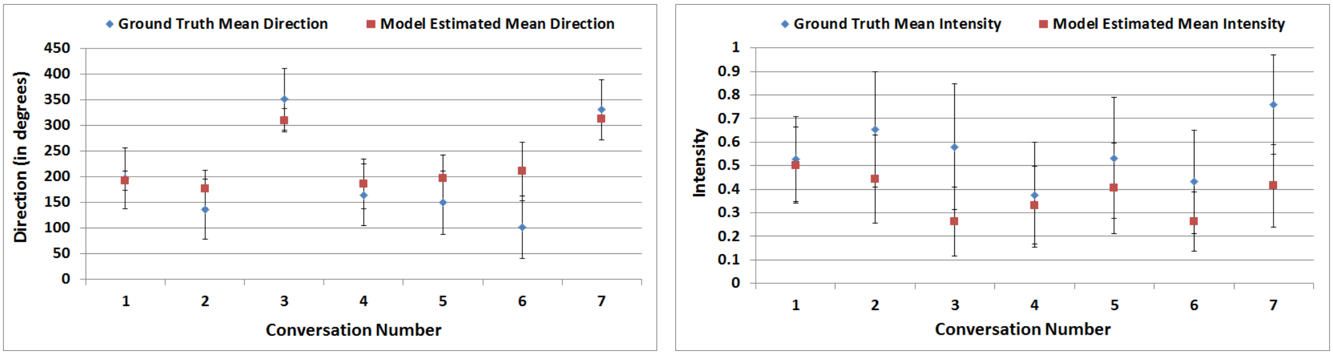
Fig. 4 shows the model estimated cdf of each of the emotions in the mixture model separately and the empirical cdf for the first dialog in the testing set. It can be seen that the model estimated cdf for anger (Fig. 4(a)) as well as that of frustration (Fig. 4(b)) fits the empirical cdf well. However, the cdfs for happiness (Fig. 4(c)), excitement (Fig. 4(d)), and sadness (Fig. 4(e)) show quite high deviation from the empirical cdf. This suggests that the empirical distribution and the model estimated distributions of anger as well as frustration are not statistically different. On the basis of this, we may say that the first conversation in the testing set is an angry/frustrated conversation.

Figs. 3 and 4 show that the proposed mixture model fits the data well, despite the underlying problems with the ground truth labelling (that is, the fact that there is only one label associated with each utterance, which lasts for many frames while the model estimates the values for each frame). Furthermore, all 'silent' frames are unlabelled in the conversations, while the model estimates the values for those frames as well. The intensity values do not fit as well as the directions because the small number of samples leads to high concentration around the mean as compared to the large number of frames in the testing set.

Fig. 5 shows the mapping of continuous emotions through time corresponding to a full continuous dialog (Ses01F_impro01) in the activation-evaluation space. The colour variation represents time, ranging from red (dark in grayscale) to yellow (light in

TABLE 1: Mean intensity of five emotions and neutral state estimated using dependent and independent models both with and without performing speaker-normalisation.

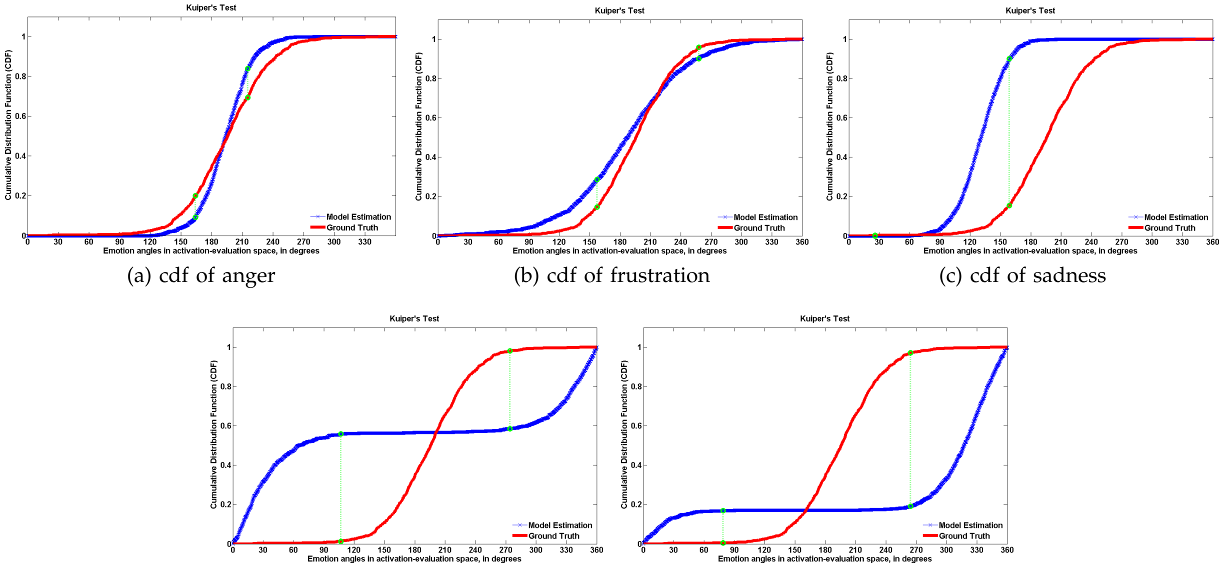
Without Normalisation						
	Neutral	Anger	Frustration	Happiness	Excitement	Sadness
Dependent Model	0.3095	0.5167	0.6066	0.4168	0.6152	0.4968
Independent Model	0.1286	0.2698	0.1809	0.2704	0.2278	0.1663
With Reference-based Speaker Normalisation						
Dependent Model	0.3095	0.5167	0.6066	0.4168	0.6152	0.4968
Independent Model	0.1693	0.3325	0.2759	0.3523	0.3628	0.2414



(a)

(b)

Fig. 3: The test of fit for (a) the mean ground truth *directions* and those estimated by the mixture model (b) the mean ground truth *intensities* and those estimated by the mixture model, for each of the seven conversations in the test set. Lines mark one standard deviation.



(a) cdf of anger

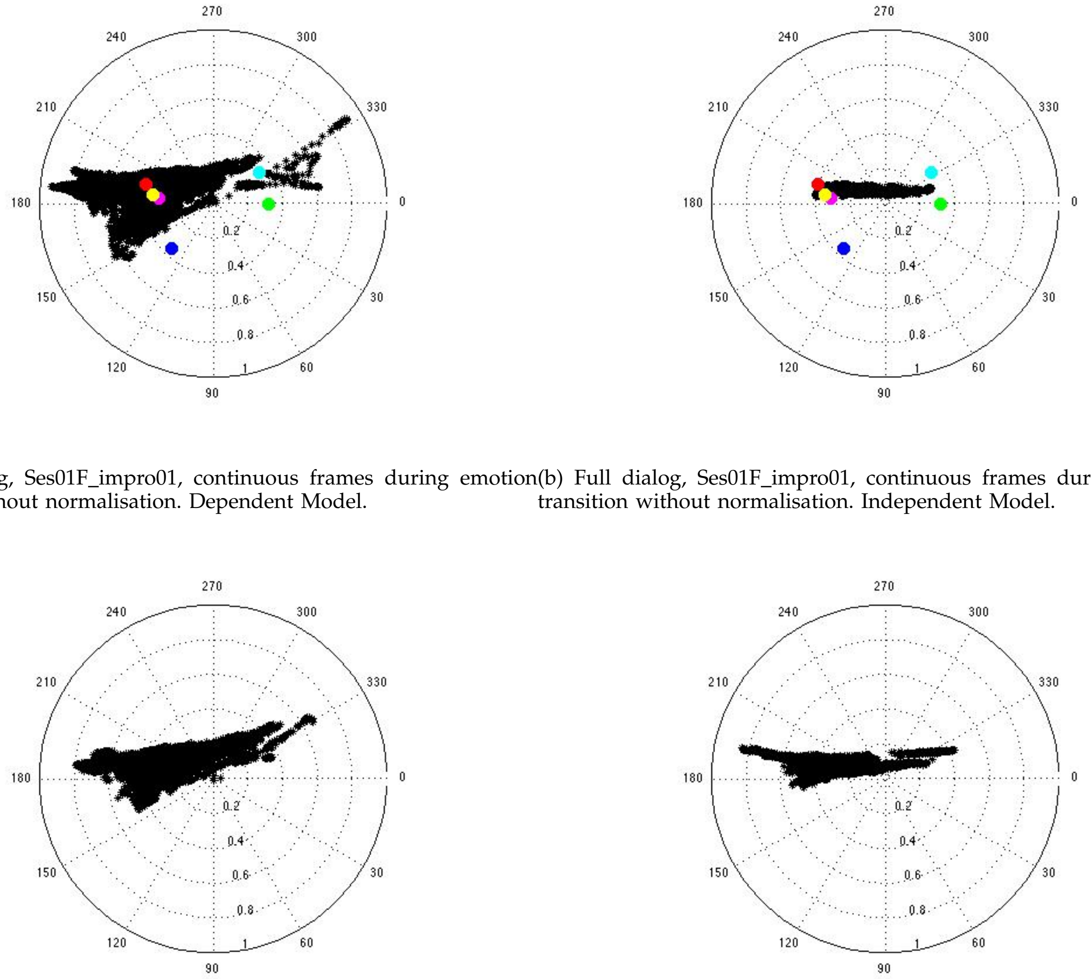
(b) cdf of frustration

(c) cdf of sadness

(d) cdf of happiness

(e) cdf of excitement

Fig. 4: Illustration of the Kuiper's statistic. Blue line is model estimated CDF of (a) anger, (b) frustration, (c) sadness, (d) happiness, and (e) excitement, red line is an empirical CDF of the first conversation in the testing set, and the green line is the Kuiper's statistic.



(a) Full dialog, Ses01F_impro01, continuous frames during emotion transition without normalisation. Dependent Model.

(b) Full dialog, Ses01F_impro01, continuous frames during emotion transition without normalisation. Independent Model.

(c) Full dialog, Ses01F_impro01, continuous frames during emotion transition with normalisation. Dependent Model.

(d) Full dialog, Ses01F_impro01, continuous frames during emotion transition with normalisation. Independent Model.

Fig. 5: With and Without normalisation, dependent versus independent model

grayscale). The dialog has 16 female turns consisting of 13835 frames. Most of the turns are labelled as angry/angry/frustrated by the three human observers, which matches the observation well. Fig. 5(a) shows mapping based on speaker-dependent model without normalisation, Fig. 5(b) shows mapping based on speaker-independent model without normalisation, Fig. 5(c) shows mapping based on speaker-dependent model with reference-based normalisation, and Fig. 5(d) shows mapping based on speaker-independent model with reference-based normalisation.

Explain Gamma

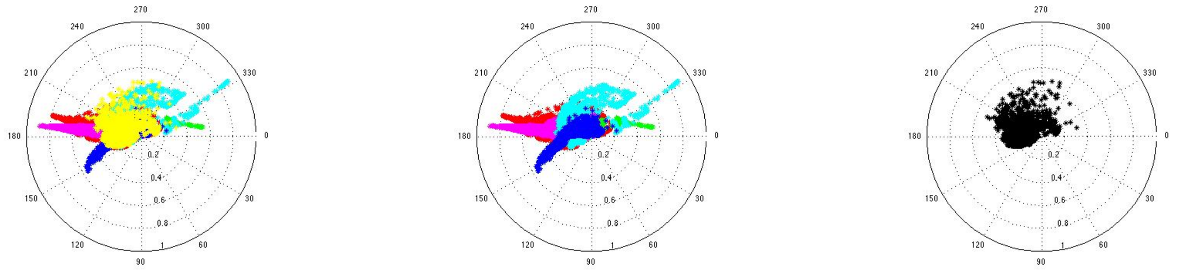
5 Analysis of Emotion Dynamics

The research related to the study of 'discrete emotion dynamics' focuses mainly on the detection of four temporal segments: neutral, where there is no sign of activation of any facial expression; the onset of a facial expression, when the muscular contraction begins and increases in intensity; the apex, which is the peak where the intensity reaches a stable level; and the offset, which is the relaxation of the muscular action back to the neutral state.

However, emotions are not discrete: they continuously change over time due to their natural progression, external stimuli, and the way the face works. Naturally, emotions fade in their intensity with time [?]: the intense anger that might have been accompanied by betrayal by a close friend might provoke a milder response when recalled after weeks or months. Similarly, the intense feeling of joy accompanied by winning a championship might provoke a milder sense of happiness when looking at the event's photos weeks or months later. Also, research shows

	Dependent Model	Independent Model
Without Normalisation	70.3503	63.2036
With Normalisation	56.1678	**78.9750**

TABLE 2: Mean accuracy dependent and independent model with and without reference-based speaker normalisation.



(a) All gamma emotions

(b) All gamma emotions excluding neutral

(c) Gamma neutral

Fig. 6: Gammas of all emotions without normalisation

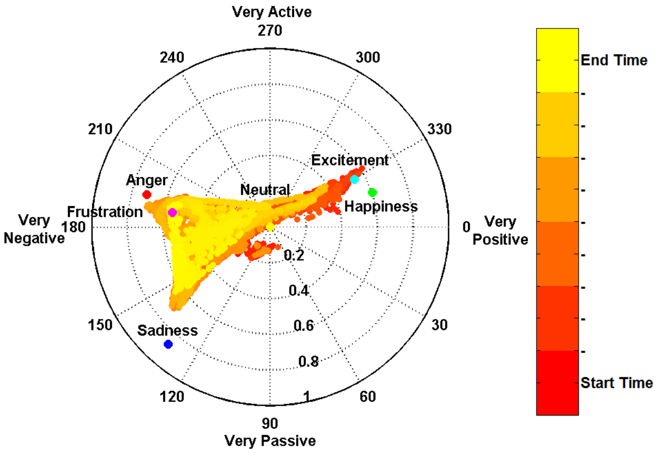


Fig. 7: The mapping of continuous emotions during a conversation (Ses01F_impro01) into the activation-evaluation space. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end).

that a noticeable change in emotions is brought about by an external stimulus, e.g., the behaviours of others, or a change in the current situation, or internal stimuli such as thoughts or memories [?].

In addition, emotional *expressions* change based on the mechanical properties of facial skin. The facial dermal tissues comprise collagen (72%) and elastin (4%) fibres which help resist deformation of tissues. Therefore, the facial tissues effected by the active muscle activity caused by emotion expressions need to relax before stretching to another form (i.e., expressing another emotion) [?]. Moreover, emotions are related to each other in a systematic manner which guides the way emotions go through transition from one state to another. For example, the transition from anger to frustration is more common than the transition from anger to happiness.

So far, none of the computational studies have focused on the temporal analysis of 'continuous emotions dynamics' in order to understand the relationships between different emotions and the paths followed by emotions while moving from one state to another (i.e., emotion trajectories).

This paper addresses these problems by taking into account continuous spontaneous expressions of emotions. It focuses on the study of complex emotions as the weighted set of basic emotions, considering their relationships with each other. In addition, this paper presents ways to study the temporal dynamics of continuous emotions in order to get insight about the emotion trajectories and the corresponding intensity variations over time.

The paths of emotional expressions are observed in the activation-evaluation space. Fig. 7 shows the mapping of continuous emotions through time corresponding to an unsegmented conversation in the activation-evaluation space (using the methods described in Section 3). The colour variation represents time, ranging from red (dark in grayscale) to yellow (light in grayscale). The figure shows some sequence information in the transition from one emotion state to another. Also, we already knew the mechanical properties of facial dermal tissues and that the emotions are related to each other in a systematic manner, as discussed in Section 1. All these facts and findings motivated us to computationally analyse the paths of emotions while transitioning from one state to another. We propose the following hypotheses about the emotion paths representation in the activation-evaluation space:

1) The paths among emotions form 'smooth' trajectories in the space.
2) If the end-point emotions are not positively correlated, then the path goes through the neutral state.
3) If the end-point emotions are positively correlated, then the path does not go through the neutral state.

Positively correlated emotions are those that lie close to each other (less than 90° divergence) in the activation-evaluation space. Uncorrelated emotions correspond to the emotions that lie at 90° divergence, while the negatively correlated emotions corresponds to the emotions which are 180° apart in the activation-evaluation space.

5.1 Evaluation

Within the following section the methods used to test the proposed hypotheses are presented, together with the results of those tests.

Hypothesis 1: The paths among emotions form 'smooth' trajectories in the space.: In order to test the first hypothesis, we first need to define 'smoothness' of an emotion trajectory. After extensive review of the literature, we could not find any standard definition. However, if we consider an emotion trajectory as a time series (sequence of values at successive time points following a non-random order), then its smoothness may be defined as a measure of its persistence with time. A random time series, e.g., Brownian motion, is not smooth, as it is not persistent with time. On the basis of this definition, we may say that if the points in the emotion space move in a predictable manner then the resulting trajectory is smooth/persistent with time.

We measure the smoothness of emotion points trajectories in the activation-evaluation space using two approaches: *first*, by measuring the time derivative of angular displacement (angular velocity) and *second*, by estimating the Hurst exponent (H) [**?**]. The time derivative of angular displacement determines the change in the angle with time; the smaller the change, the smoother the trajectory and vice-versa. It is approximated by:

$$\dot{\theta}_t = \theta_t - \theta_{t-1}$$

where $t = 2, 3, \cdots, n$ and n = total number of frames in the video. We plotted the time derivatives of angular displacement of the emotion points during each conversation in the testing set and got smaller values that imply smooth emotion trajectories [**?**].

The Hurst exponent is a statistical measure of persistence and predictability of a time series, calculated by rescaled range $\left(\frac{R_t}{S_t}\right)$ analysis, where R_t is the rescaled range and S_t is the standard deviation of the time series. The calculation of the rescaled range R_t of the time series will be described shortly. The greater the value of H $(0.5 < H < 1)$, the smoother the time series, $H = 0.5$ means a random time series. We plotted the the Hurst exponent as a measure of smoothness of emotion trajectories in the activation-evaluation space for each conversation in the testing set and got all values close to 1 [**?**]. We estimated H for the time series representing the size of 'change' between pairs of consecutive points in the space as a function of valence and activation. Suppose X_t denotes the time series where $t = 2, 3, \cdots, n$ and n = total number of frames in the video. The size of 'change' for each frame is calculated as the Euclidean distance between the time derivative of valence $(\dot{V_t})$ and the time derivative of activation $(\dot{A_t})$ in the activation-evaluation space:

$$\dot{V_t} = V_t - V_{t-1}$$
$$\dot{A_t} = A_t - A_{t-1}$$

where the t index represents the t^{th} element of the time series.

$$\text{size of change} = \sqrt{\dot{V_t}^2 + \dot{A_t}^2} \tag{14}$$

The rescaled range R_t of time series (X_t) is calculated by:

1) Calculate the mean-centred time series $Y_t = X_t - \mu$, where μ is the mean of the time series.
2) Calculate the cumulative sum of Y_t,
$$Z_t = \sum_{i=1}^{t} Y_i$$
3) Calculate the rescaled range R_t of time series,
$$R_t = \max(Z_1, Z_2, \cdots, Z_t) - \min(Z_1, Z_2, \cdots, Z_t)$$

S_t is the standard deviation of the time series. The ratio $\frac{R_t}{S_t}$ scales as a power law with time so that:

$$H = \frac{\log\left(\left(\frac{R}{S}\right)_t\right) - c}{\log(t)}$$

where c is a constant and the slope of the regression line (R/S versus t in log-log axes) approximates the Hurst exponent. Figure 8 shows the linear regression model fitted to the R/S analysis of all emotion trajectories in the testing set.

We calculated the size of 'change' between two consecutive emotion points in the activation-evaluation space for each conversation in the testing set separately and found that the size of 'change' (which is the motion within 120^{th} of a second) is mostly very small. However, in a few cases the size of change is bigger. In order to find the reason behind these intensity jumps (those beyond the first standard deviation), we monitored those paths of trajectories and compared them with the original videos in the dataset. We noticed that the bigger size of change in the trajectories are false positives due to closing the eyes. We had tried to avoid this by removing the eyelid markers, but still the closing of eyes is captured by the muscles around the eyes, especially those near the eyebrows. Fig. 9 shows some of the false positives

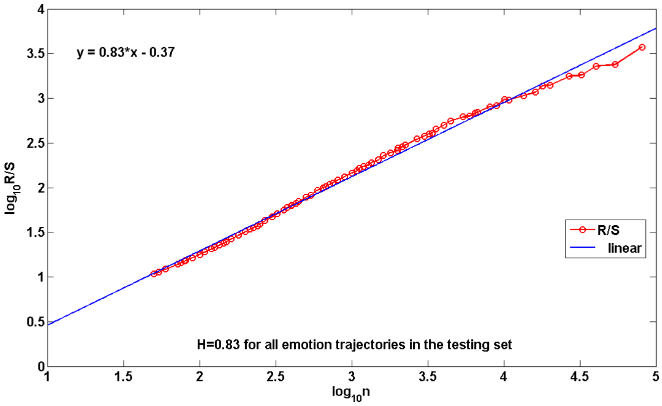


Fig. 8: A linear regression model fitted to R/S analysis for all emotion trajectories in the testing set.

during transition from excitement to frustration. As lowering eyelids/eyebrows is one of the expressions of sadness, the outliers in the emotion space lies in the direction of sadness. It should be noted that the size of 'change' refers to the size of displacement between the two consecutive points in the activation-evaluations space, not the displacement of markers on the face.

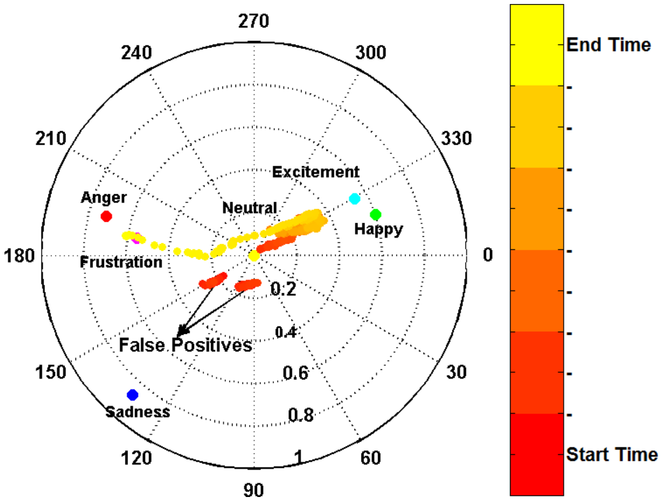


Fig. 9: The false positives appear during transition from neutral to excited. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end).

Hypotheses 2 and 3: Transitions between uncorrelated or negatively correlated emotions need to pass through the neutral state, while transitions between positively correlated emotions do not.: According to differential emotion theory, each discrete emotion is related to certain other discrete emotions with a distinct pattern [?]. In the activation-evaluation space, emotions lie along particular angles on the basis of their similarity measures.

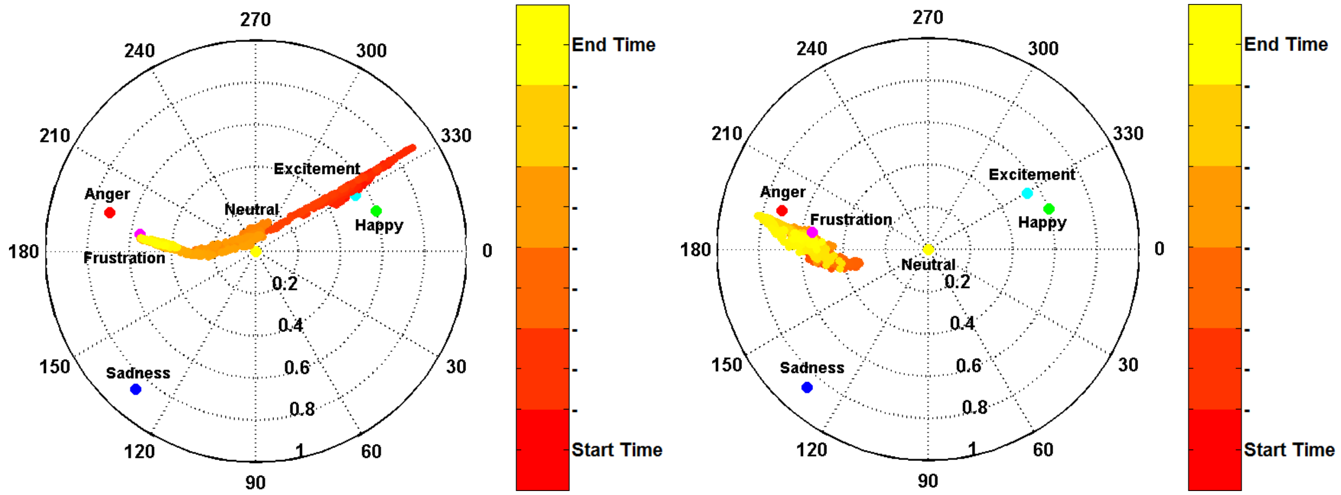
While looking at continuous trajectories in the activation-evaluation space (e.g., Fig. 7), we noticed some patterns of emotion transitions. We fitted regression lines to the trajectories among six emotions (neutral, happiness, excitement, anger, frustration, and sadness). There are 15 possible symmetric paths among the six emotions, but the regression analysis shows that actually only seven paths are represented for all possible transitions among these emotions. For instance, the *smiling* expression (getting happy) shows a linear relationship of valence and activation between neutral and happiness. The onset of smiling occurs at neutral, reaches apex at some intensity of happiness and returns to neutral for offset. In the case of *laughing*, which is another expression of happiness, the same pattern repeats several times depends on its duration and intensity.

By observing the trajectories between two uncorrelated and negatively correlated emotional states, we found that the transition between these states tends to pass through the neutral state. The intensity of the current emotion must decrease to neutral (shown as linear motion along a radial line) before the intensity of the next emotion increases. Fig. 10(a) shows a trajectory followed by the transition from excitement to frustration. However, the positively correlated emotions (such as anger and frustration, as well as happiness and excitement) may move from one state to another with slight change in intensity and angle simultaneously, as shown in Fig. 10(b). These findings are also supported by the mechanical properties of facial dermal tissues [?]. Under low stress (transition between two positively correlated emotions), dermal tissue applies low resistance to stretch as the collagen fibres uncoil in the direction of the strain. However, under high stress conditions (transition between two negatively or uncorrelated emotions), the elastin fibres behave like elastic springs to return the collagen fibres to their original no-stress condition. According to these properties, to express a very different emotion the facial muscles have to pass through a 'no-stress' condition.

Emotion Transition	Linear (R^2)	Quadratic (R^2)	Cubic (R^2)
Neutral-Anger	0.9131	0.9297	0.9139
Neutral-Frustration	0.9642	0.9854	0.9924
Neutral-Happiness	0.9223	0.9323	0.9418
Neutral-Excitement	0.9374	0.9383	0.9388
Neutral-Sadness	0.902	0.9144	0.9164
Anger-Frustration	0.0922	0.6494	0.6497
Happiness-Excitement	0.0457	0.5914	0.6083

TABLE 3: Coefficient of determination (R^2) of linear, quadratic, and cubic polynomial regression models fitted to the seven symmetric paths of emotion transitions into the activation-evaluation space.

Table 3 shows a comparison among the coefficient of determination (R^2) of linear, quadratic, and cubic polynomial regression models fitted to each of the seven symmetric paths (i.e., between neutral and happiness, neutral and excitement, neutral and anger,



(a) Continuous transition between excitement and frustration, which are negatively correlated emotions

(b) Continuous transition between anger and frustration, which are positively correlated emotions

Fig. 10: Transitions between negatively correlated emotions tend to pass through the neutral state, while transitions between positively correlated emotions do not.

neutral and frustration, neutral and sadness, anger and frustration, and happiness and excitement). For the paths from neutral to any of the five emotions, there is no significant difference between the (R^2) values of linear, quadratic and cubic regression, which implies that a linear model may be used to fit these paths. For the path between neighbouring emotions such as anger to frustration as well as happiness to excitement, the quadratic regression is significantly better than the linear regression. Moreover, there is no significant difference between quadratic and cubic regressions, which implies that quadratic model may be used to fit the trajectories between the neighbouring emotions onto the activation-evaluation space.

In the activation-evaluation space, the travel along the emotion flows/trajectories is a matter of intensity change and the angle change. As already discussed, the emotion trajectories follow 'common' paths, which in turn suggests that there exists some relationship between the intensity change and angle change through time. In order to analyse this relationship, we plot the polar coordinates ($\dot{r}$: changing intensity, $\dot{\theta}$: changing emotion) of continuous points in the space during emotion transitions. Fig. 11 consists of four subplots showing r and θ respectively. In these plots, the three horizontal lines represent the mean (μ) and mean $\pm$ 1 standard deviation σ. The third and fourth subplots show change versus no-change using a binary plot by applying the following rules to the time derivatives $(\dot{r}_t, \dot{\theta}_t)$ of (r, θ) respectively (where the t index represents the t^{th} element of the time series):

$$f(\dot{\theta}_t) = \begin{cases} 0, & \text{if } \mu_{\dot{\theta}_t} - \sigma_{\dot{\theta}_t} < \dot{\theta}_t < \mu_{\dot{\theta}_t} + \sigma_{\dot{\theta}_t} \\ 1, & \text{otherwise} \end{cases}$$

$$f(\dot{r}_t) = \begin{cases} 0, & \text{if } \mu_{\dot{r}_t} - 2\sigma_{\dot{r}_t} < \dot{r}_t < \mu_{\dot{r}_t} + 2\sigma_{\dot{r}_t} \\ 1, & \text{otherwise} \end{cases}$$

Fig. 11 shows that there is a relationship between angle change and intensity change such that whenever there is a large 'change' in angle (according to the given rules), the intensity decreases.

6 Discussion

This paper has presented our approach to map facial points to an appropriate emotion-space to analyse the paths of emotions while moving from one state to another. We develop a shape model for both speaker-dependent and speaker-independent discrete emotion recognition. It give us the distance of each frame from five emotion states and the neutral state. We perform speaker-reference based normalisation to minimise the effect of speaker-face variability. We compared the results of our system with the majority label out of more than three labels assigned by three human evaluators and the speaker-independent model gives highest recognition accuracy after reference-based normalisation.

The research in psychology has shown that in everyday interactions people exhibit non-basic, intermediate, or complex emotional states which are related to each other in a systematic manner. Based on this, it is not appropriate to assign a single independent categorical label to complex emotions which are mixtures of more than one emotion category. Also, sometimes the difference between two or more emotions is so subtle that it becomes difficult to finely differentiate between them. Due to this ambiguity, human evaluators assigned more than one categorical label to each turn.

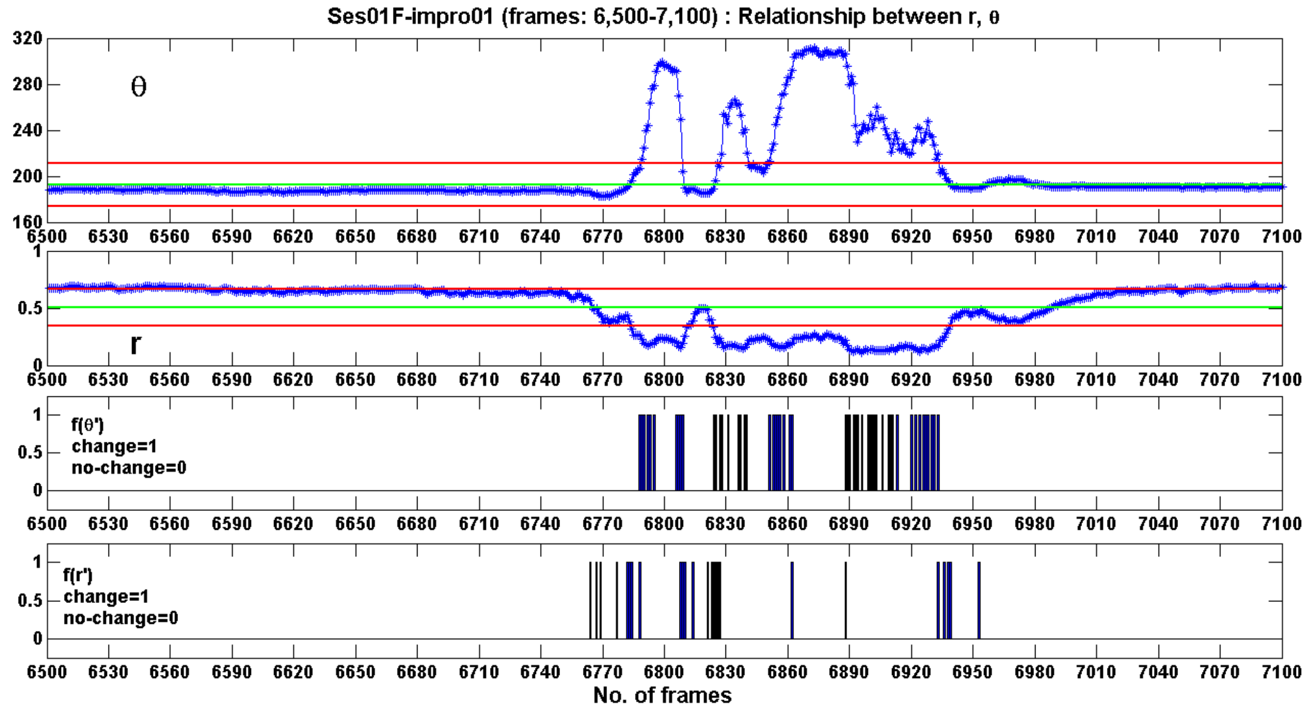


Fig. 11: Ses01F_impro01, continuous frames during emotion transition from anger to happiness. The temporal relationship between intensity change and angle change can be seen in the frames from 6750 to 6950. For explanation, see the text.

Motivated by these thoughts, in this paper we address the problem of spontaneous complex emotion recognition by mapping them into the emotion space. Instead of classifying the emotions to one emotion category based on the best-matched criteria, we add the effect of all basic emotions for the recognition of complex emotions.

Using the von Mises mixture model, we presented a technique for the recognition and representation of complex emotions in the activation-evaluation space. The proposed method is based on the psychological assumption that complex emotions are comprised of mixtures of basic emotions. There is still debate among psychologists on the number of basic emotions and which emotions should be considered as basic, and of the six emotions that we have considered two (frustration and excitement) are candidate basic emotions. However, the proposed mixture model is quite flexible and can be applied to any set of basic emotions. We estimated the degree of similarity of each test frame to each of the basic emotions and project them into the activation-evaluation space using the von Mises mixture model.

Using von Mises mixture model, each continuous conversation is mapped into the activation-evaluation space frame by frame. However, we know that emotions are related to each other in a systematic manner which guides the way emotions go through transition from one state to another [?]. In order to understand the relationships among different emotions and the paths followed by emotions while moving from one state to another (i.e., emotion trajectories), we extended this work to the computational analysis of continuous emotion trajectories in the activation-evaluation space 5. Emotions vary in intensity, flow, persistence with time, and their relationships with other emotions. By analysing the emotion dynamics through time, we try to seek the answers about the 'common' paths between emotions, the smoothness of emotion trajectories, and how we travel along emotion flows. The computational analysis of emotion dynamics may be helpful for better understanding of emotion trajectories as well as in the development of more flexible models for emotion recognition, representation, and synthesis.

Section 5 presented an analysis of emotion trajectories in the activation-evaluation space based on shape models of facial points. On the basis of trajectory-level analysis, we evaluated some hypotheses related to the smoothness of emotion trajectories, and the 'common' paths among emotions based on their correlation.

By measuring the size of 'change' between two consecutive frames, we found that the emotions move in a continuous flow, which implies that there are no sudden jumps within the trajectories. Further, we measured the smoothness of continuous emotion trajectories on the basis of the time derivative of angular displacement and the estimated Hurst exponent, which suggests that the emotion trajectories are smooth and persistent with time.

By visualising the emotion trajectories, we observed that there are 15 possible symmetric paths among the six emotions in the activation-evaluation space. To test it, we fitted regression lines to the trajectories and found that there are actually only 9 symmetric paths to travel among these six emotions. We showed that a linear model fits well to the trajectories between neutral and any of the other five emotions. The trajectories between uncorrelated and negatively correlated emotions cannot be fitted with one linear model, however, two linear models are better fitted than one quadratic or cubic regression model. A quadratic regression model fits well to the trajectories between positively correlated emotions.

By analysing the relationship between the change in angle and change in intensity, we may conclude that the transition between negatively correlated or uncorrelated emotions causes a decrease in intensity, while the transition between positively correlated emotions may occur with a slight change in intensity and angle simultaneously.

The presented analysis might be used and extended in several directions, such as examining the 'abnormal' paths of emotions, which might give some cues about underlying deception, or some illness. The mapping of continuous trajectories to the activation-evaluation space might be a useful tool to build emotional conversation agents displaying realistic emotions and going through smooth emotion transitions.

Acknowledgments

This work was supported by funds from the Higher Education Commission (HEC) Pakistan. The authors would like to thank Massey University for providing support throughout the research.

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982.

[3] C. E. Izard, *The Psychology of Emotions*. Springer, 1991.

[4] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Actions*. Palo Alto: CA: Consulting Psychologists Press, 1978.

[5] C.-L. Huang and Y.-M. Huang, "Facial expression recognition using model-based feature extraction and action parameters classification," *Journal of Visual Communication and Image Representation*, vol. 8, no. 3, pp. 278–290, 1997.

[6] T. F. Cootes and C. J. Taylor, "Combining point distribution models with shape models based on finite element analysis," *Image and Vision Computing*, vol. 13, no. 5, pp. 403–409, 1995.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Training models of shape from sets of examples," in *BMVC92*. Springer, 1992, pp. 9–18.

[8] S. Marsland, *Machine Learning: An Algorithmic Perspective*. Chapman and Hall CRC, 2009.

[9] H. Hong, H. Neven, and C. V. der Malsburg, "Online facial expression recognition based on personalized galleries," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 354–359.

[10] L. Wiskott, *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*. Deutsch, 1995.

[11] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 1. IEEE, 2004, pp. 592–597.

[12] J. R. Movellan, "Tutorial on gabor filters," *Open Source Document*, 2002.

[13] P. Michel and R. Kaliouby, "Real time facial expression recognition in video using support vector machines," in *IEEE International Conference on Multimodal Interfaces (ICMI)*, 2003, pp. 258–264.

[14] P. Martins and J. Batista, "Identity and expression recognition on low dimensional manifolds," in *IEEE International Conference on Image Processing*. IEEE, 2009, pp. 3341–3344.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Computer Vision-ECCV*. Springer, 1998, pp. 484–498.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, pp. 585–591, 2001.

[17] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image and Vision Computing*, vol. 24, no. 6, pp. 605–614, 2006.

[18] A. Bansal, S. Chaudhary, and S. D. Roy, "A novel LDA and HMM-based technique for emotion recognition from facial expressions," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Springer, 2013, pp. 19–26.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[20] D. Sontag and D. Roy, "Complexity of inference in latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2011, pp. 1008–1016.

[21] E. Costantini, F. Pianesi, and M. Prete, "Recognising emotions in human and synthetic faces: The role of the upper and lower parts of the face," in *Proceedings of the 10th International Conference on Intelligent User Interfaces*. ACM, 2005, pp. 20–27.

[22] J. N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face." *Journal of Personality and Social Psychology*, vol. 37, no. 11, pp. 2049–2058, 1979.

[23] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, 1976.

[24] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Journal of Image and Vision Computing*, vol. 27, no. 12, pp. 1741–1844, 2009.

[25] M. Heller and V. Haynal, "Depression and suicide faces," *What the Face Reveals*, pp. 398–407, 1997.

[26] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. WW Norton & Company, 2009.

[27] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[28] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 36, no. 1, pp. 35–43, 1999.

[29] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[30] B. Braathen, M. S. Bartlett, G. Littlewort, E. Smith, and J. R. Movellan, "An approach to automatic recognition of spontaneous facial actions," in *Proceedings Fifth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2002, pp. 360–365.

[31] M. H. Mahoor, S. Cadavid, D. S.Messinger, and J. F. Cohn, "A framework for automated measurement of the intensity of non-posed facial action units," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 1 & 2, 2009, pp. 833–839.

[32] I. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation - The standard, Implementations and Applications*. John Wiley and Sons, 2002.

[33] M. Pantic and L. J. M. Rothkrantz, "An expert system for recognition of facial actions and their intensity," in *Seventeenth National Conference on Artificial Intelligence (AAAI-2001) / Twelfth Innovative Applications of Artificial Intelligence Conference*, 2000, pp. 1026–1033.

[34] G. Zhou, Y. Zhan, and J. Zhang, "Facial expression recognition based on selective feature extraction," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 2. IEEE, 2006, pp. 412–417.

[35] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Journal of Neural Networks*, vol. 18, pp. 423–435, 2005.

[36] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 4, no. 42, pp. 335–359, 2008.

[37] M. Kipp, "ANVIL - A generic annotation tool for multimodal dialogue," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, 2001, pp. 1367–1370.

[38] A. Ortony and T. J. Turner, "What's basic about basic emotions," *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[39] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 4, pp. 5–60, 1999.

[40] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, April 1936, pp. 49–55.

[41] A. Hakim, S. R. Marsland, and H. W. Guesgen, "Statistical modelling of complex emotions using mixture of von mises distributions," in *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

Ayesha Hakim Ayesha Hakim has a BS(Hons.) in Computer Science from Bahauddin Zakariya University, Multan, Pakistan, which she completed in 2006. She received a gold medal for reaching the first position in the university and obtaining the highest CGPA record. She worked as Assistant Professor of IT in the same university. Along with the lectureship, she spent time working as a software engineer in a multinational company. On the basis of her excellant academic background, the Higher Education Commission of Pakistan granted her the Indigenous Scholarship to complete a doctrate in computer science. She joined Massey University in July 2008, where she completed her PhD in 2014. Her research interests include affective computing, human-computer interaction, mobile computing, and context-aware systems.

Stephen Marsland Stephen Marsland has a BA(Hons) in Mathematics from the University of Oxford and a PhD in "Self-Organisation and Novelty Detection" from the University of Manchester, which he completed in 2002. Since then he has spent time at the Santa Fe Institute, the University of Bremen and the University of Manchester, where he was a lecturer in computer science and a researcher in the divison of Imaging Science and Biomedical Engineering. He moved to Massey University in 2004, and was awarded an Early Career research award there in 2005. He is currently a professor and the postgraduate director of the School of Engineering and Advanced Technology at Massey University.

Hans W. Guesgen Hans Guesgen holds a diploma in computer science and mathematics of the University of Bonn, a doctorate in computer science of the University of Kaiserslautern, and a higher doctorate (Habilitation) in computer science of the University of Hamburg, Germany. He worked as a research scientist at the GMD in Sankt Augustin from 1983 to 1992, primarily in the area of artificial intelligence and expert systems. In 1992 he joined the Computer Science Department of the University of Auckland, where for almost fifteen years he led the AI research group. In 2007, he was appointed Chair of Computer Science in the School of Engineering and Advanced Technology at Massey University.