Estimation of the Temporal Dynamics of Posed and Spontaneous Facial Expression Formation using LLE

Jane Reilly Delannoy and John B. McDonald
Computer Vision and Imaging Laboratory
Department of Computer Science, National University of Ireland Maynooth
jreilly@cs.nuim.ie

Abstract

When analysing facial expressions, it is not only the final expression itself, but also its formation that plays an important role when attempting to decipher its meaning. Currently in research there are two techniques for describing the dynamics of facial expression; quantitative and temporal based analysis. Quantitative-based techniques attempt to determine the amplitude of the expression in terms of intensity levels, where the levels correspond to some measure of the extent to which the expression is present on the face. Temporal-based techniques split the expression into three temporal phases (onset-apex-offset). In this paper we focus on the temporal aspects of facial expression formation, describing our research into applying a non-linear manifold extraction technique for modelling these temporal phases. We present initial results of our technique for modelling the temporal aspects of both posed and spontaneous facial expressions.

1. Introduction

Over the past number of decades computer vision researchers have created systems specifically for the automatic analysis of facial expressions. The most successful of these approaches draw on the tools of behavioral science, where many different standards and methodologies for encoding facial expressions were developed (see [7] for a review). The *Facial Action Coding System* (FACS), created by Ekman *et al.* in 1978, is the most comprehensive of these standards and is widely used in research. The FACS provides an unambiguous quantitative means of describing all possible movements of the face in terms of 46 *Action Units* (AUs) [6], 27 of which can be directly linked to facial expression and emotion.

With a view to developing ubiquitous computing solutions, in recent years, researchers have pushed the boundaries of human-computer interaction to a stage where current techniques are beginning to reach a level of maturity necessary for their application in real-world applications and interfaces. Towards this aim, systems which react to ones facial expressions are starting to come on stream, such as that developed by Whitehill *et al.* [9], where they created a system which enables the user to fast forward and rewind video based on predetermined facial expressions.

Initial research into the classification of facial expressions focused on identifying the six prototypical expressions. However, over the past number of years significant technological advances have enabled researchers to develop systems which automatically classify the AUs involved in facial expressions in real time [2] with a classification rate of 93% on posed datasets, and 75% on spontaneous datasets.

Research by behavioural scientists has shown that it is not only the expression itself, but also its dynamics that are important when attempting to decipher its meaning [4, 1]. The dynamics of facial expression can be defined as the intensity of the facial movement coupled with the timing of its formation. Ekman *et al.* suggest that the dynamics of facial expression provides unique information about emotion that is not available in static images [7].

Currently in research there are two main approaches for describing the dynamics of facial expression, quantitative analysis of the degree of intensity of the expression, and analysis of the temporal segments or phases of the expression sequence. Quantitative based techniques attempt to determine the extent to which the expression is present on the face in terms of levels of intensity such as the five levels of intensity as defined by Ekman *et al.*in the FACS [7], or the three level model of intensity defined by Delannoy *et al.* [5]. Temporal based analysis of expression sequences, divides the expression sequence in to three temporal segments (onset, apex and offset).

In this paper we focus on analysing the temporal aspects of facial expression formation, presenting initial results from our LLE-based technique for modelling the tem-

978-0-7695-3796-2/09 $26.00 © 2009 IEEE
DOI 10.1109/IMVIP.2009.32

IEEE computer society

poral aspects of both posed and spontaneous facial expressions.

The remainder of this paper is structured as follows, in Section 2 we discuss the dynamics of posed and spontaneous facial expression, and provide an overview of both the quantitative and temporal methods of describing the dynamics of facial expressions. Details of our proposed technique for analysing the of temporal aspects of the dynamics of facial expression are provided in Section 3. In Section 4 we detail our experiments. Finally in Section 5 we discuss our experimental results and provide future direction.

2 Extracting dynamical information from facial expression sequences

Despite the fact that facial expressions can be either subtle or pronounced in their appearance, and fleeting or sustained in their duration, most of the studies to date have focused on investigating displays of extreme posed expressions rather than the more natural spontaneous expressions.

Posed facial expressions are generally captured by asking subjects to perform specific facial actions or expressions. They are usually captured under artificial conditions, i.e. the subject is facing the camera under good lighting, with limited head movement, and the expressions are usually exaggerated. Spontaneous facial expressions are more representative of what happens in the real world, typically occurring under less controlled circumstances. With spontaneous expression data, subjects may not necessarily be facing the camera, the image size may be smaller, there will undoubtedly be a greater degree of head movement, and the facial expressions portrayed are often less exaggerated.

The dynamics of posed expressions can not be taken as representative of what would happen during natural displays of emotions, similar to how individual words spoken on command would differ from the natural flow of conversation. Consequently, when analysing the dynamics of facial expressions, one must realise that while the final image in a posed sequence will be the requested facial expression, the sequence as a whole may not allow for the accurate modelling of the interplay between the different movements that make up the facial expression during its natural formation.

As one would expect, when studying the dynamics of facial expression formation, the key information is extracted by analysing the interplay between the facial features which caused the expression to appear on the face. Therefore, although one can create systems which are capable of modelling the dynamics of posed expressions, the extension of these systems to deal with spontaneous facial expressions is a non trivial task.

From the perspective of labelling or annotating the dynamics of facial expression foramtion, regardless of the type of data used, the dynamics of facial expressions are described using either quantitative or temporal labels. While these labels can be used along with prior knowledge to differentiate between posed and spontaneous expressions, the act of labelling the data itself does not take these factors into account.

2.1 Quantitative Dynamics

As mentioned earlier, the FACS provides and unambiguous means of describing all movements of the face in terms of Action Units (AUs). Where an AU describes the movements of one or more muscles in the face that causes an atomic change in the faces appearance. However facial expressions do not always appear to the same degree and for this reason the FACS also includes intensity levels for the AUs. There are five intensities in total ranging from intensity A, where a trace change in appearance occurs, to intensity E, where an extreme appearance change occurs.

Although the FACS provides a good basis for AU and intensity coding of facial images by human observers, the way in which the AU and intensity codes have been defined does not easily translate into a computational test. One of the issues with the accurate application of these labels is that the intensity levels are defined as a series of descriptions, and as such are open to a certain element of subjectivity in their application. Also the five levels are not evenly distributed across the evolution of an expression, for example intensity C occurs for a longer period than intensity A during the formation of a given AU.

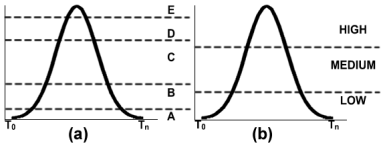

Figure 1. *Illustration of the FACS 5 levels of expression intensity (a) and the three level model of expression intensity (b). The y-axis represents intensity, and the x-axis represents increasing time. Where* $\mathbf{T}_o$ *is the 1st in an expression sequence, and* $\mathbf{T}_n$ *is the final frame in an expression sequence.*

As a means of mitigating these problems, a three level model of facial expression intensity was introduced by Delannoy *et al.* [5]. The main advantage that the three level model has over the FACS five level model, is that it divides the intensity spectrum according to three intuitive levels, of low, medium and high intensities, rather than the complicated labelling of the FACS. Experiments have shown that this results in a more robust repeatable estimation of intensity across the expression sequence [5].

Regardless of which approach you use to quantitatively

analyse the dynamics of facial expression, a certain level of error is to be expected as during the formation of a facial expression. This is due to the continuous deformation of the face over time, which makes it difficult to determine an absolute boundary between the various intesnity levels.

2.2 Temporal Dynamics

The onset-apex-offset method, as the name suggests, divides the expression into three temporal phases. This can effectively be represented as a two stage model of expression intensity, as shown in Fig. 2 (a). In the *onset phase* the initial movements during the expression formation take place. In the *apex phase* the expression peaks. The *offset phase* is effectively a mirror of the onset phase in that the expression fades back to neutral, often merging into the onset phase of another expression.

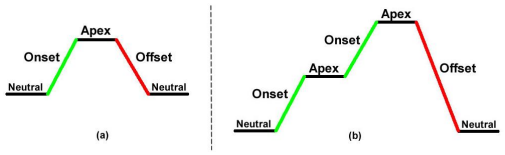

Figure 2. *Assuming that the first frame in the sequence is the neutral expression, once facial movement commences the onset state starts, when the expression plateaus the expression enters the apex state. From the apex there are two possibilities, (a) the facial expression may begin to decrease in intensity hence the offset phase commences or (b) the expression may increase in intensity, hence the onset state commences.*

One of the common misconceptions with temporal dynamics of facial expressions is the assumption that the expression reaches an extreme level of expression intensity during the apex of an expression. While this can often be the case, it is quite possible for multiple apexes to be present during the formation of natural facial expressions as shown in Fig 2 (b), or indeed when analysing subtle facial expressions, the expression may apex at low intensities. It is for this reason that it is difficult to ascertain when the apex of the expression has occurred without analyzing the entire expression sequence. Pantic *et al.* mitigate this problem by assuming that facial movement plateaus during the neutral and apex phases of the expression sequence.

Automatic approaches towards analysing the dynamics of posed facial expression data have been developed by Pantic *et al.* where they use rule based techniques to explicitly analyse the temporal dynamics of facial expressions in terms of these three phases using 15 fudicial facial points which are tracked throughout the expression sequence [14]. More recently Koelstra and Pantic have improved on these results by incorporating appearance based information in their experimentation [10].

Initial works which focused on differentiating between posed and spontaneous facial expression using temporal analysis have been presented, such as that of Littlewort *et al.* where they developed a technique which differentiated between real and posed pain, achieving a 72% accuracy in a two-way forced choice [12]. While this is an impressive result, it should be noted that the posed expressions used in this experiment were exaggerated versions of the steriotypical pain expression, while spontaneous facial expressions which were much more subtle and varied in their appearance.

In a study on the key indicators of driver fatigue, Vural *et al.* recorded subjects playing driving video games for prolonged periods. Here they used information relating to the timing and intensity of the appearance of the facial signals of tiredness, such as blink rate, eye closure and head tilt to determine whether a driver was in a drowsy or alert state with 90% accuracy [20].

Results such as these have demonstrated that the dynamics of expression formation can be used to accurately differentiate between real and posed facial expressions. In this paper we present our manifold based technique for the temporal analysis of posed and spontaneous facial expressions.

3 Experimental structure

In this section we appraise the efficacy of our *Locally Linear Embedding* (LLE) based technique for estimating the temporal dynamics of posed and spontaneous facial expression data. Firstly we provide details of the datasets used in our experiments. We then discuss the techniques which we use to extract the features relating to the various AUs from our datasets. Following on from this we provide details of the LLE algorithm.

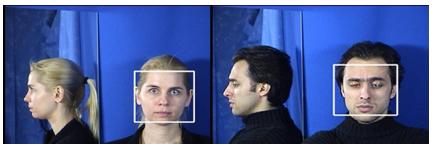

Figure 3. *Displays sample frames extracted from video sequences from the MMI facial expression dataset. The features which we automatically detect for each frame (face outline, and center of left and right eyes) are also shown*

3.1 Datasets

As input to our experiments on posed facial expressions, we used video sequences from the *MMI FACS coded facial expression dataset* (MMI database) [13], an example of which is shown in Fig.3. The MMI database contains approximately 2894 sequences of various individuals performing AUs both in isolation and in combination with other AUs. The sequences contain the normal temporal dynamics of an expression formation, in that they go from neutral - onset - apex - offset - neutral, although some sequences contain multiple apexes.

Due to the limited availability of spontaneous facial expression datasets, for our experiments on spontaneous expressions we used frames from an in-house database of individuals performing *Irish Sign Language* (ISL). This dataset contains video sequences of a native ISL signer performing various signs to another ISL signer who was located beside the video camera. During the video sequences there is a large degree of head rotation and unlike the posed expression data the interframe movements were more pronounced. An example of frames extracted from the sign language sequences are shown in Fig. 4.

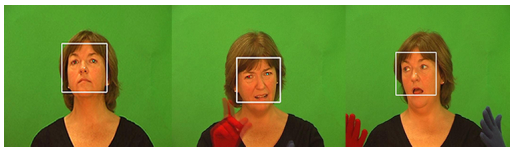

Figure 4. *Displays sample frames extracted from our in-house ISL video sequences. The features which we automatically detect for each frame (face outline, and center of left and right eyes) are also shown*

3.2 Feature Extraction

The computational analysis of the dynamics of facial expression using textural data is not a straightforward problem. The first task is the accurate detection of the facial features involved with the various AUs. In our research we detect the face and eyes in each frame of the video sequences using the OpenCV implementation of the Viola-Jones face detector [19], along with open source Haar cascades for face and eyes [17]. The features which we extract from each frame in the video sequences, shown in Fig. 3 and Fig. 4.

Once we have detected the central points for the left and right eye, we calculate *Regions Of Interest* (ROI) for the various AUs. This is done by combining our knowledge of the FACS AU descriptions along with studies on facial geometry. Fig. 5 illustrates the ROI's for the upper face AUs. Since there are six principal AUs which alter the appearance

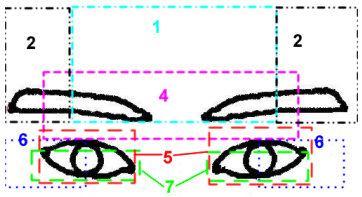

Figure 5. *This figure illustrates the regions of interest for each AU in the upper face. By using ROI's we were able to focus on the specific areas of the face which is altered by the AUs*

of the upper face (AU1,AU2,AU4,AU5,AU6,AU7), we have defined six ROIs for this region, one for each AU. The appearance changes due to combinations of AUs occurs in the overlapping sections. For example, AU1 raises the inner brow, while AU4 lowers the brow, the apperance changes due to the combination of AU1 and AU4 occur in the intersection of these two ROIs.

Once we have the ROI for the AUs in question defined, we further reduce the variance in the dataset by performing image differencing on the image sequences, whereby we subtract the neutral image in the sequence from each subsequent image in the sequence.

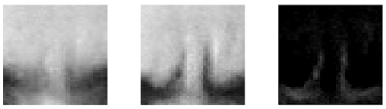

Figure 6. *From left to right this figure displays region of interest for AU4 for the the neutral brows, extreme and difference image*

By reducing the variance in our dataset by focusing on the regions of interest for particular AUs, and performing image differencing, the input to the LLE algorithm is a series of vectors which describe the deformations which occurred in the texture of that specific facial region. An example of this can be seen from Fig. 6 where (iii) is the difference between the frames (i) and (ii).

3.3 Local Linear Embedding (LLE)

The LLE algorithm was introduced by Saul and Roweis in 2000 as an unsupervised learning algorithm that computes low dimensional, neighbourhood preserving embeddings of high dimensional data [18].

It is based on simple geometric intuitions, where the algorithm computes a low dimensional representation of the

high-dimensional imput data, with the property that points which were located within the same neighbourhood in the input data, are similarly co-located with respect to one another in the low dimensional embedding. Since its introduction in 2000, many extensions and adjustments to this core algorithm have been proposed, ranging from Robust-LLE [3, 8] to supervised [16, 11] and semi-supervised versions of LLE [15], however in this paper we are using the original algorithm as defined in [18].

4 Temporal analysis of facial expressions

As a result of our feature extraction, the variance in our dataset caused by altering facial expression is enhanced, whilst the variances caused by altering identity are suppressed. This enables us to exploit the neighbourhood preserving property of the LLE algorithm to estimate the underlying manifold of increasing expression intensity as the expression formed. This resulting lower dimensional manifold facilitates the estimation of the temporal aspects of facial expression formation on a frame-by-frame basis.

For both our posed and spontaneous datasets, prior to experimentation, two trained FACS coders analyse the sequences and apply temporal labels on a frame by frame basis. These labels form our ground truth, and form the basis for evaluating the results output by our technique. It is important to note that due to space limitations, in this paper we do not deal with the issue of facial expression classification, therefore we use expression sequences which have been previously FACS coded as input to our experiments.

4.1 Automatically applying temporal labels to posed and spontaneous facial expressions

In this experiment we used data from the MMI facial expression database. We chose a sample of 20 sequences in total which contain the entire temporal pattern of expression formation in that at the start of the video the subject's expression is neutral, following on from this they perform an expression and then return to neutral. This enabled us to establish a complete picture of the temporal dynamics of posed facial expressions

For our posed facial expression experiment, we extracted our features as described in Section 3.2, and applied LLE to each expression sequence individually. From studying the first dimension of the LLE output along with these labels, we observed that during the formation of an expression, the sequences followed the same general structure.

That is, the neutral expressions were projected into the leftmost region of the LLE space and the extreme or most intense expression were located in the rightmost region of the LLE space. Also we found that the frames which we had manually labelled as belonging to the onset phase traversed this space from left to right (increase in x-value), similarly the frames in the offset phase traversed this space from right to left (decrease in x-value).

Using this information we developed a model based on an examination of the x-values within this 1-dimensional LLE space. If the value increased by more than a threshold value, the onset phase label was applied. If the value remained constant then depending upon the previous state, it was labeled as being in the neutral or apex state. If the value decreased by more than the threshold value then the offset state was applied. An example of how our technique performed when compared with the temporal labels applied by the human FACS coders is illustrated in Fig. 7.

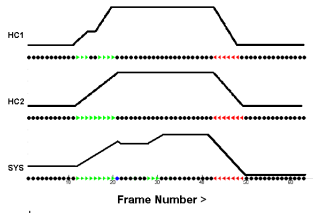

Figure 7. *Comparison between the temporal labels as output by our technique (SYS) and those of two independent coders (HC1, HC2) for a particular expression sequence. The green right arrow represents the onset phase, the red left arrow represents the offset phase, and the star denotes the frames during which the expression plateaued*

Overall when applied to expression sequences from the MMI posed facial expression database, our system had an agreement rate of 88% with our first human coder, and 93% with our second human coder, giving an overall mean agreement rate of 90% when the labels of the two human coders are combined.

While this is a promising result, our next step was to determine if this approach was suitable for estimating the temporal dynamics of spontaneous facial expressions using our in-house dataset. Automatically applying temporal labels to spontaneous facial expression data is a more complex problem as the onset of one expression may occur during the offset of another expression. Also multiple onsets are possible within one video sequence and the expression sequence does not always follow the prototypical path from neutral to onset to apex to offset to neutral.

It is quite possible for an expression to move between states in an irregular fashion during its formation, this is illustrated in Fig. 8. When viewed alongside Fig. 7, the differences in the temporal patterns of posed and spontaneous facial expressions are clearly apparent. Consquently these results suggest that the temporal patterns of expression for-

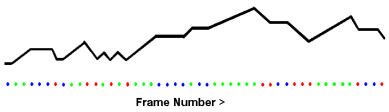

Figure 8. *Illustration of the temporal labels output from our technique for a particular expression sequence from our in-house ISL dataset. The peaks and hollows represent the apex/neutral phases, the upward lines represent the onset phase, and the downward arrows represent the offset phase*

mation could potentially be used to differentiate between posed and spontaneous facial expressions.

5. Conclusions

In this paper we have demonstrated the efficacy of Locally Linear Embedding, when combined with the feature extraction techniques described in Section 3.2, for estimating the temporal phases of posed facial expression sequences, achieving a mean agreement rate of 90% with human coders.

Using our in-house Irish Sign Language dataset, we have demonstrated that this technique provides the necessary basis for the automatic estimation of the temporal phases of spontaneous facial expressions.

From the results of our experiments we have shown that our technique for analysing the temporal patterns of facial movement within our Regions of Interest can be used to differentiate between posed and spontaneous facial expressions. Future work will involve applying this technique to larger datasets of posed and spontaneous facial expressions and also to other regions of the face.

References

[1] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 2005.

[2] M. Bartlett, G. Littlewort, C. Frank, M.G.and Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia 1(6) p. 22-35.*, 2006.

[3] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. *IEEE International Workshop on Analsysi and Modelling of Faces and Gestures*, pages 25–35, 2003.

[4] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. *Proceedings of Intel. Conf. On Multimedia and Expo, 2001.*, 2002.

[5] J. R. Delannoy and J. B. McDonald. Automatic estimation of the dynamics of facial expression using a three-level model of intensity. *In proceedings of the IEEE conference on Automatic Face and Gesture Recognition*, 2008.

[6] P. Ekman, W. Friesen, and J. Hager. Facial action coding system. *Consulting Psychologists Press*, 1978.

[7] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System Manual*, 2002.

[8] A. Hadid and M. Pietikaanien. Efficient locally linear embeddings of imperfect manifolds. *Proceedings of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany*, pages 188–201, 2003.

[9] J. M. J. Whitehill, M. Bartlett. Automatic facial expression recognition for intelligent tutoring systems. *Proceedings of IEEE Computer Vision and Pattern Recognition Conference 2008*, 2008.

[10] S. Koelstra and M. Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. *Proceedings of IEEE conference on Automatic Face and Gesture Recognition*, 2008.

[11] O. Kouropteva, O. Okun, A. Hadid, M. Soriano, S. Marcos, and M. PietikAainen. Beyond locally linear embedding algorithm. Technical report, Department of Electrical and Information Engerring, University of Oulu, Oulu, Finland, MVG-01-2002, 2002.

[12] G. Littlewort, M. Bartlett, and K. Lee. Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proc. International Conference on Multimodal Interfaces, Nagoya, Japan.*, 2007.

[13] M. Pantic and I. Patras. Mmi au-coded facial expression database. Technical report, Twente University, 2005.

[14] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *SMC-B*, 36(2):433–449, April 2006.

[15] D. D. Ridder and R. P. W. Duin. Locally linear embedding for classification. Technical report, Pattern Recognition Group, Department of Imaging Science and Technology, Delft University of Technology, Delft, The Netherlands, PH-2002-01, 2002.

[16] D. D. Ridder, O. Kouropteva, O. Okun, M. PietikAanien, and R. P. W. Duin. Supervised locally linear embedding. *ICANN 2003*, pages 333–341, 2003.

[17] M. C. Santana, O. Dniz-Surez, L. Antn-Canals, and J. Lorenzo-Navarro. Face and facial feature detection evaluation: Performance evaluation of public domain haar detectors for face and facial feature detection. *VISAPP*, 2008.

[18] L. K. Saul and S. T. Roweis. An introduction to locally linear embedding. http://www.cs.toronto.edu/ roweis/lle/publications.html, 2001.

[19] P. Viola and M. Jones. Rapid object detection using a boosed cascade of simple features. *IEEE CVPR*, 2001.

[20] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Drowsy driver detection through facial movement analysis. In *In Proc ICCV*, 2007.