Automatic Estimation of the Dynamics of Facial Expression using a Three-Level Model of Intensity

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

Abstract

Facial expressions and their associated dynamics play an important role in human communication. The dynamics of facial expressions can be defined as the intensity and timing of their constituent components as they form. However, estimating the dynamics of facial expressions is a non trivial task. The majority of automatic approaches to characterising intensity use a two-level model (also known as onset-apex-offset). However the FACS specifies five intensity levels for each AU. In this paper we evaluate the efficacy of Local Linear Embedding as a means of estimating the intensity of facial expression. This is done using both the full five level FACS model, and a simplified three level model. We have found that using the FACS intensity scoring results in a considerable overlap between the estimated intensities. Using a three level model enables us to classify the intensities with significantly greater degree of accuracy.

1. Introduction

Since the importance of facial expressions was first established in 1872 [9], many studies have been carried out attempting to interpret their meaning. 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 behavioural science, where many different techniques for encoding facial expressions were developed (see [12] for a comprehensive review). The *Facial Action Coding System* (FACS), created by Ekman and Friesen, 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 movements of the face in terms of 46 *Action Units* (AUs) [10].

Within the field of facial expression analysis there has been a significant amount of research investigating the six prototypical expressions (anger, fear, sadness, joy, surprise, and disgust). However, in everyday life, while these primary expressions occur frequently, when analysing human interaction and conversation, researchers have found that displays of emotion or intention are more often communicated by small subtle changes in the face's appearance [1].

From a computer vision perspective, initial research into the classification of facial expressions focused on identifying the six prototypical expressions. However, in more recent years computer vision researchers have concentrated upon classifying the individual movements or AUs that make up the facial expressions. Perhaps the most substantial work in this area has been conducted by Bartlett *et al*. Bartlett *et al*. proposed a technique which combines Gabor wavelets and SVMs to classify AUs with 93.3% accuracy [3]. Again in [16], Littlewort and Bartlett propose a similar technique which classifies AUs with 97% accuracy.

Recent research has shown that it is not only the expression itself, but also its dynamics that are important when attempting to decipher its meaning [1, 3, 6, 8, 14]. The dynamics of facial expression can be defined as the intensity of the facial movement coupled with the timing of their formation. Ekman *et al*. suggest that the dynamics of facial expression provides unique information about emotion that is not available in static images [11].

In this paper we provide an overview of the main approaches currently used in research for the analysis of the dynamics of facial expression. We also detail our proposed technique for modelling the dynamics of facial expression formation. Here, once the AUs present in a facial expression sequence have been identified, our technique then classifies the dynamics of the expression in terms of our simplified three stage model of increasing intensity.

The remainder of this paper will be structured as follows. In Section 2 we discuss the theory behind the analysis of the dynamics of facial expressions. We appraise contemporary research investigating the dynamics of facial expression from a computer vision perspective in Section 3. In Section 4 we provide some background information on the techniques and methodologies which we have used in our

978-1-4244-2154-1/08/$25.00

approach. Following on from this, in Section 5 we demonstrate the success of our technique for modelling the dynamics of facial expression, discussing some experiments and work to date. The research presented in this paper builds on our previous works presented in [18, 19] and [20].

2. Dynamics of facial expression

According to Ambadar *et al*., few investigators have examined the impact of dynamics in deciphering faces. These studies were largely unsuccessful due to their reliance on extreme facial expressions. Ambadar *et al*. also highlighted the fact that facial expressions are frequently subtle. They found that subtle expressions which were not identifiable in individual images suddenly became apparent when viewed in a video sequence [1].

There is a growing trend in psychological research which argues that the dynamics of facial expression play a critical role in the interpretation of the observed behaviour. Zheng *et al*., state that an expression sequence often contains multiple expressions of different intensities sequentially, due to the evolution of the subject's emotion over time [25].

2.1. Posed vs Spontaneous Facial Expressions

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 static 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 conditions, 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 will not allow for the accurate modelling of the interplay between the different movements that make up the facial expression during its natural formation.

Recently published research has shown that the dynamics of facial expression formation can be used to distinguish between posed and spontaneous expression of emotion. For example, Littlewort *et al*. developed a technique which differentiated between real and posed pain, achieving a 72% accuracy in a two-way forced choice [15]. Vural *et al*. used information relating to the timing and intensity of the appearance of the facial signals of tiredness, such as blink rate, eye closure and yawn to determine whether a driver was in a drowsy or alert state with 90% accuracy [24].

3. Capturing dynamical information from facial expression sequences

Within the field of computer vision, two main approaches are currently used to describe the dynamics of facial expression formation. In this section we provide details of these two techniques, illustrating how our technique differs from these approaches.

3.1. Onset-Apex-Offset Phases of facial expression formation

The onset-apex-offset method, as the name suggests, divides the expression into three temporal phases. This can effectively be represented as a two level model of expression intensity, shown in Figure 1 (a). In the *onset phase* the initial movements during the expression formation take place, in the *apex phase* the expression peaks and the *offset phase* is a mirror of the onset phase in that the expression fades back to neutral, often merging into the onset phase of another expression. One of the main issues concerning the application of the onset-apex-offset model to the analysis of the intensity of facial expression formation, is that the apex phase does not equate to extreme intensity, rather it indicates when the expression is most intense for a given expression sequence.

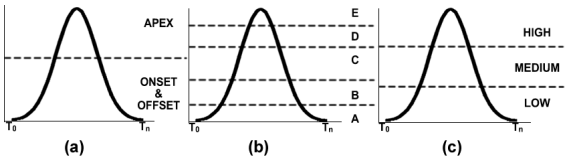

Figure 1. *Example of how the onset-apex-offset (a), the FACS (b) and our technique (c) divide up the expression into two, five or three levels of intensity respectively. In these figures, the y-axis represents intensity, and the x-axis represents increasing time. Where* $\mathbf{T}_o$ *is the first frame in an expression sequence, and* $\mathbf{T}_n$ *represents the final frame in an expression sequence.*

In general, the onset-apex-offset phases are identified post processing by human experts as seen in [2] where the results of human observers are used as a form of validation of research results. Recent studies performed by Pantic *et al*. [17], explicitly analyse the temporal dynamics of facial

expressions in terms of these three phases using rule based encoding.

3.2. Facial Action Coding System Intensity Scores

As mentioned earlier, the FACS provides an unambiguous, quantitative means of describing all movements of the face in terms of Action Units (AUs). An AU describes the movement of one or more muscles in the face that causes an atomic change in the face's appearance. However AUs do not always occur with the same intensity 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. In Figure 1 (b), we have illustrated how the FACS divides the evolution of an expression into 5 levels of intensity. However, these 5 intensity 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. The effect that varying intensity of an expression has on the appearance of the face is shown in Figure 2.

Figure 2. *Example of the effect of varying the intensity of an expression. from left to right the intensities are: Neutral, intensity A, C and E*

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. The reason for this is that the FACS is an appearance based technique with the AUs being defined as a series of descriptions. The FACS intensity coding guidelines are also description based, and as a result are somewhat subjective. Hence special effort is required to establish and maintain acceptable levels of reliability. Sayette *et al*. suggest that the reliability of intensity coding may be problematic and state that further work is needed [23].

The main contribution of this paper lies in the development of a simplified model of the dynamics of facial expression formation as illustrated in Figure 1 (c). Our proposed technique differs from previous works as it is repeatable across different regions of the face, and provides a more representative model of the underlying dynamics of facial expression formation. More specifically our technique differs from the onset-apex-offset model as our technique provides for the automatic identification of the intensity of facial expression formation on a frame-by-frame basis, whereas the onset-apex-offset model is mainly concerned with identifying the onset, apex and offset phases across the expression sequence as a whole. Details of how our technique and hypothesis are provided in Section 5.

4. Proposed Methodology

In this section we provide background information on the techniques which we implement for modelling the dynamics of facial expression formation. Once an expression has been classified (our classification technique builds on the works of Reilly *et al*. [19]), we extract information regarding the dynamics of the expression formation by first projecting shapes of individuals portraying the specific expressions into a lower dimensional *Locally Linear Embedding* (LLE) space in order to capture the underlying manifold of that expression as it forms. Where the manifold of facial expression, refers to the concept that facial expressions and their formations define a smooth underlying manifold in low dimensional space [5]. Once we have extracted the manifold of the expression formation we classify the particular intensity of the expression using *Support Vector Machines* (SVMs). We validate our results using *Receiver Operating Characteristic* (ROC) curve analysis. In this section we provide some background details on the different techniques we implement in our research.

4.1. Local Linear Embedding

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 [22].

The LLE algorithm is based on simple geometric intuitions where the algorithm attempts to compute a low dimensional embedding with the property that nearby points in the high dimensional space remain nearby and similarly co-located with respect to one another in the low dimensional space. As input, LLE takes a dataset of N real valued vectors $\mathbf{X}_i$, each of dimensionality D, sampled from some smooth underlying manifold. Provided there is sufficient data such that the manifold is well sampled, we can expect each data point and its neighbours to lie on or close to a locally linear patch of the manifold [22].

LLE takes place over three steps, where firstly the manifold is sampled, and for each sample, the *K* nearest neighbours are identified. Secondly each point $\mathbf{X}_i$ is approximated as a linear combination of its neighbours $\mathbf{X}_j$. These linear combinations are then used to construct the sparse weight matrix $\mathbf{W}_{ij}$. Reconstruction errors are then measured by the cost function $\sum \mathbf{W} = \sum_i |\mathbf{X}_i - \sum_j \mathbf{W}_{ij}\mathbf{X}_i|^2$, which adds up the squared distances between all the data points and their reconstructions.

Finally each high dimensional observation $\mathbf{X}_i$ is mapped to a low dimensional $\mathbf{Y}_i$, which best preserves the geometry of $\mathbf{X}_i$'s neighbourhood by fixing the $\mathbf{W}_{ij}$'s and minimising

$\phi\mathbf{Y} = \sum_i |\mathbf{Y}_i - \sum_j \mathbf{W}_{ij}\mathbf{Y}_i|^2$. For more details on the LLE algorithm see [22].

4.2. Support Vector Machines (SVM)

SVMs are a type of learning algorithm based upon advances in statistical learning theory [4, 21] and are based on a combination of techniques. The *kernel trick* is one of the principal ideas behind SVMs, where data is transformed into a high dimensional space making linear discriminant functions practical. SVMs also use the idea of large margin classifiers. Suppose we have a dataset $(x_1, y_1), ..., (x_m, y_m) \in \mathbf{X} \times \{\pm 1\}$ where $\mathbf{X}$ is some space from which the x_i have been sampled. We can construct a dual Lagrangian of the form $W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ which are subject to the constraints $\alpha_i \geq 0 \quad \forall i$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$ The solution to this equation is a set of α values which defines a hyperplane that is positioned in an optimal location between two classes.

5. Experiments and Results

In our research to date we have used the *Cohn-Kanade AU-Coded Facial Expression Database* (CK-database) [7]. This database contains approximately 2000 image sequences from over 200 subjects. The subjects came from a cross-cultural background and were aged approximately 18 to 30. The CK-database contains full AU coding and partial intensity coding of facial images and is the most comprehensive database currently available.

Since the CK-database is not completely FACS intensity coded, one of the first problems which we had to overcome in our research was to manually FACS intensity code the facial images ourselves. This was a non-trivial task as although an expression sequence may be labelled as containing a particular AU or AU group, these AUs may not necessarily appear at the same degree of intensity. A sample of the final images from expression sequences taken from CK-Database for AU25 is shown in Figure 3, where it is clear that these expressions are not all of the same intensity.

Figure 3. *Examples of AU25 classified mouths from the Cohn-Kanade facial expression database*

Prior to experimentation we preprocess our data by firstly aligning the data using *Generalized Procrustes Alignment* (GPA) [13]. GPA aligns two shapes with respect to position, rotation and scale by minimising the weighted sum of the squared distances between the corresponding landmark points. Following on from this we apply *Shape Differencing*, whereby the neutral expression shape of each subject is subtracted from the sample set for that subject. This enables us to effectively uncover the underlying manifold facial expression formation independent of identity.

Using this preprocessed data we classify the expression in terms of the AUs present using the techniques described by Reilly *et al*. in [19]. Once we know which AUs are present, we apply the LLE algorithm to create a one dimensional manifold, which describes the expression formation going from neutral expression through the various intensity levels to the extreme expressions.

We hypothesise that the neighbourhood preserving property of the LLE algorithm will cause the data to be clustered according to expression intensity, thereby extracting the manifold of the increasing intensity of the AUs as the expression forms. Hence the trajectory of the expression in the LLE space allows us to develop dynamical models of the expression formation in terms of the increase in expression intensity over time. The dynamical models shown in 4, were created by fitting Gaussians to our pre-labelled LLE data, where there were insufficient samples in our dataset we plotted the mean of the sample set.

5.1. Estimating the FACS Five Levels of Expression Intensity Using LLE

In this experiment we developed dynamical models of expression formation using the FACS intensity range, A - E, as shown in Figure 4 (i). Here the distributions represent the progression of AU25 from neutral state through the 5 FACS intensities, i.e. displaying the increase in AU intensity over time. However, as can be seen from Figure 4 (i), there is a significant overlap between the intensities in the mid ranges, for example intensity D covers a large portion of the axis, demonstrating the confusion between the classes.

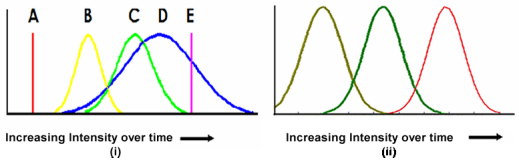

Figure 4. *Distributions for our FACS based dynamical model (i) and our simplified dynamical model (ii) for AU25, showing the Gaussians that have been fit to the 1D results of LLE algorithm which lie on the x-axis. The overlap between the distributions shows the ambiguity between intensities.*

From a practical perspective, what this means is that due to the overlap between the different intensity distributions, there will be a high false negative rate for intensity D as the data belonging to this group could get miss-classified as belonging to the other intensities. Similarly, there will also be a low true positive rate for intensity C, as effectively its data would fall under the intensity D distribution. We hypothesise that the overlap between the different intensity distributions as shown in Figure 4 (i) is due to the fact that the

FACS intensity codes are quite subjective. Differentiating between the five intensities across our dataset is a challenging problem.

5.2. Applying a three level model of intensity

Due to the problems associated with the FACS intensity coding of facial expression sequences, we began exploratory analysis looking for a better representation of facial expression intensity. From re-examining our data along with the outputs from the LLE algorithm, we observed that the data naturally clustered into three groups across the expression formation, corresponding to low, medium and high intensity displays. The results of the clustering of the dataset under the three-category labelling is shown in Figure 4 (ii).

Using this three level model, in this experiment we extract information regarding the intensity and timing of the formation of AU1+2 - *which raises the inner and outer eyebrow*, which has been previously classified. The extraction of this information provides a means for analysing the dynamics of facial expression. The input to this experiment consisted of 10 subjects from multi-cultural backgrounds, sampled from across the entire expression intensity range, 68 frames in total, labelled as being low, medium or high intensities.

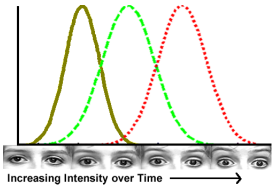

Figure 5. *Increasing intensity over time distribution for AU1+2, here we have demonstrated how our technique models the increasing intensity as the expression forms*

This experiment is split into two tasks, where firstly we apply the LLE algorithm to create a one dimensional dynamical model of the expression formation, using our simplified three stage intensity model. By exploiting the neighbourhood preserving property of LLE, in this low dimensional space our data clustered into three groups, corresponding to our three stage model of expression intensity. Using the one dimensional outputs from the LLE algorithm, we fit gaussians to the samples from each of the intensity levels. The resulting intensity distributions are shown in Figure 5, where we have also shown the affect that increasing intensity has on the appearance of the eyebrow region, by tracing the progression of the expression AU1+2 from neutral to extreme intensity.

Following on from this we divide our dataset into testing and training sets using a 10-fold cross validation strategy. In our experiments we use one-against-all *Support Vector Machine* (SVMs), as we have three groupings to classify

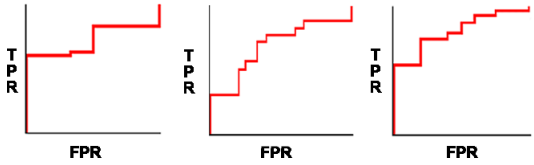

Figure 6. *ROC curve results for new 3 intensity model, from left to right are Low, Medium and High intensities*

$Intensity$	AUC	TP	FP	TN	FN
Low	0.702	4	10	41	13
Medium	0.688	34	20	9	5
High	0.793	43	16	6	3
$AverageAUC$	0.727				

Table 1. *Confusion Matrices for the three intensity levels*

we use three one-against-all SVMs. Finally we appraise our results by performing ROC analysis on the outputs of our SVMs.

The resulting ROC curves are shown in Figure 6 along with the confusion matrices in Table 1. From the confusion matrices we can see that our technique achieves an average AUC of 0.727 across the three intensity levels. The high false positives reported for the three intensities, particularly the medium intensity, can be attributed to the fact that in applying any of the models (i.e. onset-apex-offset, FACS five-levels, or our three level model) there is an inherent uncertainty in the point at which one level ends and the next begins. Hence in labelling the data there will be a consequent ambiguity at the level boundaries across the dataset.

6. Conclusion and Discussion

The accurate modelling of the dynamics of facial expression is a non-trivial task. In this paper we have proposed an alternative to the current methods for describing the dynamics of facial expressions. The solution described in this paper takes a multidisciplinary approach drawing together psychological tools, statistical models and machine learning techniques. We first build a shape model that was based on an anatomical analysis of facial expression - FACS. The FACS provided us with a universal method of analyzing facial expression and allowed for the classification of facial expressions independent of identity. In our experimental section we illustrated the subjectivity of the FACS AU intensity codes, while demonstrating the success of our simplified three stage intensity model at modelling the dynamics of facial expression in terms of Low, Medium and High intensity. Due to our small dataset we applied a 10-fold cross validation strategy, using ROC curve analysis to appraise our results, achieving an average AUC of 0.727. This is a significant result as the classification of different intensity levels is a challenging problem.

In this paper we have shown that LLE provides an effec-

tive for estimating the intensity of facial expressions. This intensity information, in conjunction with timing information, provides the necessary basis for the automated analysis of facial expression dynamics. Future work will entail applying this technique to more comprehensive datasets containing both posed and spontaneous facial expression data, incorporating a larger variety of AUs and individuals.

7. Acknowledgements

This publication has emanated from research conducted with the financial support of the Science Foundation Ireland.

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, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions., 2006.

[3] M. Bartlett, J. Movellan, G. Littlewort, B. Braathen, M. G. Frank, and T. J. Sejnowski. Towards automatic recognition of spontaneous facial actions. *In book what the face reveals*, 2003. Oxford University Press.

[4] C. Campbell. kernel methods: A survey of current techniques. *Neurocomputing*, 48:63–84, 2002.

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

[6] I. Cohen, N. Sebe, and T. L.C.A.G.and Huang. Facial expression recognition from video sequences: Temporal and static modeling, (2003).

[7] J. Cohn and Kanade. Cohn-kanade au-coded facial expression database. Technical report, Pittsburgh University, 1999.

[8] 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. *Proc. of Intel. Conf. On Multimedia and Expo, 2001.*, 2002.

[9] C. Darwin and P. Ekman. *The expression of the emotions in man and animal*. Chicago: The University of Chicago Press, 1998. 1st edition in 1872, 2nd edition in 1889, 3rd edition with additional commentry by Paul Ekman in 1998.

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

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

[12] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System Investigator's Guide*, chapter Chapter 11 FACS in Relation to Other Facial Measurement Systems. 2002.

[13] J. C. Gower. Generalised procrustes analysis. *Psychometrika*, 40:33–50, 1975.

[14] A. Hadid and M. Pietikdinen. An experimental investigation about the integration of facial dynamics in video-based face recognition. *ELCVIA*, 5(1):1–13, March 2005.

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

[16] G. Littlewort, M. S. Bartlett, I. Fasel, J. Chenu, T. Kanda, H. Ishiguro, and J. Movellan. Towards social robots: automatic evaluation of human-robot interaction by face detection and expression classification. *Advances in Neural Information Processing Systems*, 16:1563–1570, 2004.

[17] 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.

[18] J. Reilly, J. Ghent, and J. McDonald. Investigating the dyanmics of facial expression. *2nd Int. Symposium on Visual Computing*, November 2006.

[19] J. Reilly, J. Ghent, and J. McDonald. Non-linear approaches for the classification of facial expressions at varying degrees of intensity. *Irish Machine Vision and Image Processing Conf.*, September 2007.

[20] J. Reilly, J. Ghent, and J. McDonald. *Affective Computing, focus on Emotion Expression, Synthesis and Recognition*, chapter 1 Modelling, Classification and Synthesis of Facial Expressions. I-Tech, 2008.

[21] S. Rogers. *Machine Learning Techniques for Microarray Analysis*. Faculty of engineering mathematics, University of Bristol, 2004.

[22] L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Jnl. of Machine Learning Research*, 4(119), 2003.

[23] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott. *A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression*. Springer Netherlands, 2001.

[24] 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.

[25] A. Zheng. Deconstructing motion. Technical report, EECS department, U. C. Berkley, 2000.