Facial Expression Recognition in Dynamic Sequences: an Integrated Approach

Hui Fang, Neil Mac Parthaláin, Andrew J. Aubrey, Gary K.L. Tam, Rita Borgo, Paul L. Rosin, Philip W. Grant, David Marshall and Min Chen

Computer Science Department, College of Science, Swansea University,
Computer Science Department, Aberystwyth University,
Computer Science Department, Cardiff University,
E-research Centre, Oxford University

Abstract

Automatic facial expression analysis aims to analyse human facial expressions and classify them into discrete categories. Methods based on existing work are reliant on extracting information from video sequences and employ either some form of subjective thresholding of dynamic information or attempt to identify the particular individual frames in which the expected behaviour occurs. These methods are inefficient as they require either additional subjective information, tedious manual work or fail to take advantage of the information contained in the dynamic signature from facial movements for the task of expression recognition.

In this paper, a novel framework is proposed for automatic facial expression analysis which extracts salient information from video sequences but does not rely on any subjective preprocessing or additional user-supplied information to select frames with peak expressions. The experimental framework demonstrates that the proposed method outperforms static expression recognition systems in terms of recognition rate. The approach does not rely on action units (AUs) and therefore, eliminates errors which are otherwise propagated to the final result due to incorrect initial identification of AUs. The proposed framework explores a parametric space of over 300 dimensions and is tested with six state-of-the-art machine learning techniques. Such robust and extensive experimentation provides an important foundation for the assessment of the performance for future work. A further contribution of the paper is offered in the form of a user study. This was conducted in order to investigate the correlation between human cognitive systems and the proposed framework for the understanding of human emotion classification and the reliability of public databases.

Keywords: Facial expression analysis, Dynamic feature extraction and visualisation.

1. Introduction

Facial expression analysis has long been a research area of great interest. Indeed, work beginning as early as the nineteenth century [1] demonstrated that the analysis of facial expressions was of significance. The work in [2] was the first to formalise six different expressions that contained distinctive facial content. These six expressions were summarised as typical emotional displays of: *happiness*, *sadness*, *fear*, *disgust*, *surprise* and *anger*, and are now commonly known as the *basic emotions*. Until recently, the task of facial expression analysis has been a

topic of research primarily associated with the field of psychology and much on the subject has been published in this area. However, interest broadened with the publication of the work in [3] which presented a preliminary investigation of the task of automatic facial expression analysis from a sequence of images. More recently, automatic facial expression analysis has attracted much attention particularly in the field of computer science. Some of the reasons for this are due to the advancements in related research sub-areas such as face detection [4], tracking and recognition [5], as well as new developments in the area of machine learning such as feature extraction, and supervised learning[6, 7].

Much of the recent work on facial expression analysis tended to focus on ways of capturing the 'moment' or the point in time-series data (termed: *static expression recognition*) at which a particular facial expression begins to occur and when it ends. Previous approaches have mainly concentrated on attempting to capture expressions through either action units (AU) [8, 9] or from discrete frame extraction techniques [10]. All of these methods require either manual selection in order to determine where the particular behaviour occurs or the subjective imposition of thresholds. This means that any classification is highly dependent on the subjective information in the form of a threshold or other human-derived knowledge.

The approach proposed in this paper is formulated in order to tackle the aforementioned problems and to improve the performance of facial expression recognition by exploring dynamic signals. It offers a number of advantages over existing approaches: (a) the system does not require manual specification of the frame which shows peak expression; (b) the system uses the *dynamic information* of the facial features extracted from video sequences and outperforms techniques based on static images; and (c) it does not rely on the voting from groups of frames, where errors made earlier in the process are propagated leading to incorrect classification(s).

In addition to these advantages, a novel experimental evaluation presented in this paper offers a number of different perspectives for the task of facial expression analysis. For the learning of the expressions, six state-of-the-art machine learning methods are employed. Furthermore, an investigation of those sequences which are consistently mis-classified by the automatic methods is presented. This then forms the basis for a user study, which along with the use of visualisation tools offer an insight into the consistency of human perception and machine vision.

In summary, the contributions of the work are highlighted as follows:

- A novel automatic framework for the recognition of facial expressions using the dynamics of the sequences. Specific contributions include
  - The use of a group-wise registration algorithm to improve the robustness of tracking performance;
  - Construction of a parametric space of over 300 dimensions to represent the dynamics of facial expressions;
  - The use of six state-of-the-art machine learning methods for the automatic recognition task;
  - An objective comparison between the proposed system (which utilises dynamic information) and systems which utilise static apex images.
- Investigation of the correlation between human perception and machine vision for human emotion recognition.
  - The use of a visualisation technique for the analysis and initial understanding of facial feature data, and also for identifying outliers and noise in the data;

– An intuitive user study to investigate the correlation between human perception and machine vision on facial expression recognition, and to assess the quality of a public dataset.

The remainder of this paper is structured as follows. Section. 2 presents the background material for automatic facial expression analysis and provides an overview of current approaches. Section. 3 describes the proposed approach (salient facial point tracking and feature extraction methods, and construction of dynamic signal parametric space) along with the automatic learning methods. Section. 4 details the evaluation framework that is employed as well as the experimental setup and user survey. Finally, Section. 5 concludes the paper along with some suggestions for further development.

2. Background

A system for automatic facial analysis may include many different aspects. Two of the most common are: (i) the automatic detection and classification of facial expressions - an area where much work has been carried out in the past [11, 12], (ii) realistic facial expression synthesis in computer graphics [13], which is useful for studying the perception of expressions and also realistic computer animation; and (iii) expression analysis, important for affect recognition [14].

Typical facial expression recognition systems aim to classify an input facial image or video sequence into one of the six basic emotions mentioned previously. Facial expressions are formed through the movement of facial muscles, resulting in dynamic facial features such as the deformation of eyebrows, eyes, mouth and skin. Such changes can be captured and used in order to classify a given facial expression. In broad terms, there two approaches a) Facial Action Unit (AU) based techniques and b) content-based (non-AU) techniques; summarised in Section 2.1 and 2.2 respectively.

2.1. Action Unit based expression recognition

The Facial Action Coding System (FACS) [2] is the most widely used method for describing the previously described facial movements. It defines 46 different action units (AUs) for the classification of non-rigid facial movements. This system forms the basis for many expression recognition systems [15, 16, 17, 18].

In [19], several approaches that classify expressions are compared based on action unit classification accuracies. Some of these include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), Gabor filters and optical flow. It is claimed that by utilising local spatial features, better performance for expression analysis can be achieved. However, the use of techniques such as PCA destroy the underlying semantics of the local features making it more difficult to humanly interpret the results.

The work in [16] proposes the use of a rule-based system to learn facial actions by tracking salient points. Fifteen landmarks are tracked using a colour-based observation model via a particle filter algorithm applied to profile-view face images. A rule-based system is then implemented, by measuring the displacements of these salient points, in order to classify the sequences into discrete action units.

The relationship between action units using a dynamic Bayesian network is explored in [17]. The implicit assumption of this work is that the model is capable of representing the relationship amongst all AUs. Furthermore, it is claimed that AUs with weak intensity can be inferred robustly using other high-intensity AUs.

In more recent work [20], a system for emotion detection is proposed based on dynamic geometric features for AU activation detection which is then used within a hybrid SVM-HMM framework for emotion detection. The authors provide a robust analysis of their system and test the accuracy of its components on the MMI and *Cohn-Kanade* databases. However, emotion recognition performance is assessed using only the *Cohn-Kanade* database [21], so it is difficult to assess the generalisability of the approach.

2.2. Expression recognition without Action Units

For those methods which are not based on AUs, the two most common techniques for expression recognition utilise either static images that represent the apex of the expression [22] or the temporal facial dynamics [23].

In [24], grid nodes are tracked using a *Kanade-Lucas-Tomasi* tracker and the displacements of these nodes are extracted as features for training a Support Vector Machine (SVM) in order to classify the six basic expressions. This work however only extracts geometric features after tracking.

Rather than utilising geometric features, the work in [22] implements a recognition system based on texture features called Local Binary Patterns (LBP). A boosting algorithm is then used to select the active features from an LBP histogram before being passed to an SVM classifier. In [25], LBP is extended to volume LBP (VLBP) where temporal information is also exploited. Once the features have been obtained, a nearest-neighbour classifier learner is then used for classification.

An expression recognition system for video sequences is presented in [26]. The authors use several classifier learners, such as a Naïve Bayes, Tree-Augmented-Naïve Bayes and Hidden Markov models, to classify the expressions. This is carried out using a tracker system termed *Piecewise Bézier Volume Deformation*, which extracts parameters that reflect the facial deformations.

2.3. Discussion and Contributions

One particular commonly-held view is that middle-level interpretation of facial behaviour (AU recognition) can bridge the gap between low-level features and the high-level semantics of facial expressions [18]. However, a particular drawback of AU based expression recognition is the added level of AU classification prior to carrying out any expression recognition. Errors at the AU classification stage will be propagated to the expression recognition stage, leading to decreased accuracy. The argument for the use of dynamic data over static images (or the apex of the dynamics) is two-fold. Firstly, the use of static images means that the apex of the expression must first be extracted manually. This is usually straightforward for time-series data, however the data is still restricted to a single point in time, and this step must be carried out as part of a pre-processing step. Secondly, the use of temporal dynamics has proven to be more effective, and is a key factor in distinguishing between *posed* and *spontaneous* expressions [22], [14], [27].

The framework proposed in this paper focuses on a non-AU based facial expression recognition technique. It is instead based on the *dynamics* of the facial expression sequences, and is fully automatic, when compared with existing work. A majority of the non-AU based techniques are often restricted to a relatively small number of pre-defined features and typically a single machine learning technique. However, in this paper we explore various dynamic feature representations and several state-of-the-art machine learning techniques. Thus, the work is much more extensive and comprehensive than previous studies. In addition, (compared with existing

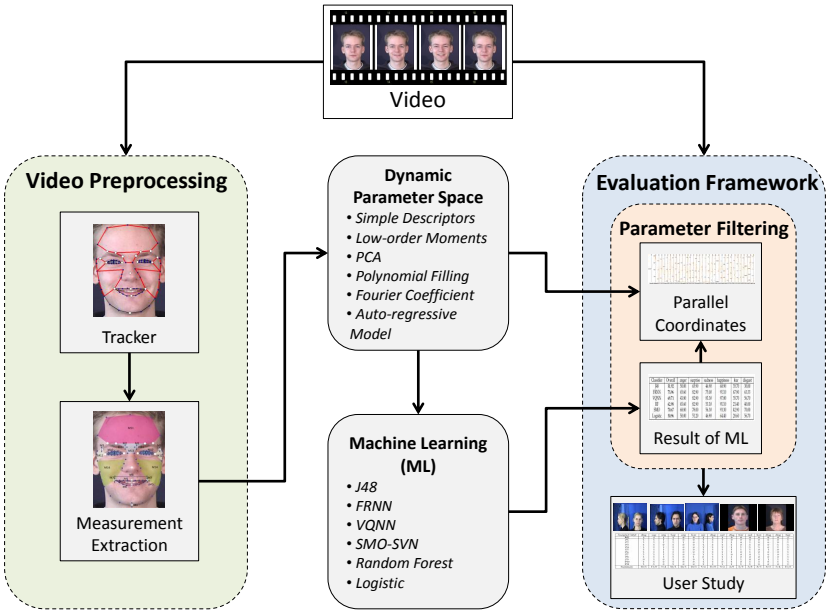


Figure 1: Proposed system for integrated facial expression analysis

techniques), it also offers a unique perspective in the form of a user study. This allows for the investigation of correlation of human perception and machine vision to be analysed.

3. Dynamic Expression Recognition Methodology

The approach proposed here integrates machine learning methods, parallel coordinates and human reasoning (in the form of a user study), in order to achieve a better understanding of the perception of dynamic changes in facial expression. An illustration of the framework is shown in Figure 1. In the following sections, the system and its components are described in detail.

3.1. Facial tracking and feature extraction

When an image sequence is presented to a facial expression recognition system, it is necessary to detect the facial regions as a preliminary pre-processing step. There are several methods which can be used to achieve this task. One of the most popular (and that which is also used in this work) is the so-called *Viola-Jones* face detector [4]. Having located the face, the next step (for both static and dynamic data), is to extract the facial features. One common approach in this respect is to *landmark* key facial points (e.g., eyes, lips, etc.) and use these to obtain the features. These landmarks can then be used to align faces in static or dynamic data and thus eliminate the effects of scaling and rotation. By tracking these points throughout a video sequence it is possible to capture the deformations (i.e. motion features) and use them for the task of expression analysis.

Most traditional trackers are template based algorithms [26, 16, 24]. These methods typically treat the first frame or neutral face image as a template and the remaining images are 'warped' to this template. Parametric deformable models which encode expected variations in shape and appearance [28, 29, 30] are extensions of the template based methods. Global shape models and appearance models, local texture models (or combinations thereof) can be used as prior knowledge in order to limit the search space. Furthermore, machine learning methods, such as linear regression [28] or graphical models [30], can be also used to locate the optimal landmarks.

Although the parametric deformable models are the most popular methods for localising landmarks, Groupwise Registration (GR) is more suitable for the proposed work. GR overcomes the limitations of a linear combination of bases and captures those subtle non-linear movements produced by different expressions. Moreover, GR has successfully been applied for facial sequences in the work in [31] and [32]. Some common elements of GR are shared with traditional registration frameworks, however GR outperforms traditional registration methods because it can obtain typical characteristics through a whole set of images rather than relying on a single template image. Moreover, it provides a dense pixel correspondence over the entire image set. In the work proposed here, piecewise affine deformation fields are used to warp the landmarks defined in Figure 2 to each face image after dense correspondences between sets of images are built based on GR. This is necessary because the deformations around those features which contain rich texture information, are more robust to image noise and the smoothing terms in the registration step.

Geometric movements, such as landmark displacements and curvature changes of facial components, play an important role in distinguishing between the expression changes in human cognitive systems. Therefore, both landmark displacements and some semantically meaningful measurements such as changes in: *eye-lid curvature*, *lip curvature*, *eye size*, etc., are extracted for the task of expression recognition. Point displacement can be represented by a set of dense points [33] or a set of sparse points [34, 16]. For the approach proposed here, the tracking algorithm is based on dense grid deformations which improve robustness when compared with tracker systems which utilise a set of sparse points. This is due to strong spatial smoothness constraints. At the same time, a sparse landmarks warper overcomes the displacement noise caused by the unpredictable changes of wrinkles from which dense optical flow algorithms typically suffer.

Once the landmarks have been tracked, the facial feature dynamics can be extracted. As discussed in Section 2, an expression comprises of several AUs. This information can be exploited to generate a list of features and associated measurements to describe each expression. The facial map in Figure 2 and the related measurements in Tables 1 and 2 describe the features and measurements used in this paper. The geometric features can be extracted from the dynamics of single or multiple points, it may even be useful to extract the curvature of features such as the upper and lower lip.

Dynamic texture changes are indispensable elements for capturing the characteristics of facial expressions. In this work, Gabor filter response energy values contained in four regions are obtained as texture features for learning expressions. These are: *cheek region*, *eye brow region*, *outer eye corner wrinkle* (often referred to as crows feet) and *forehead region*(s). Figure 2 shows the landmarks (white markers: reference points or points for getting displacements; dark blue markers: points defining regions and curves), geometric features (dark blue lines) and texture regions (coloured patches) used for each video sequence.

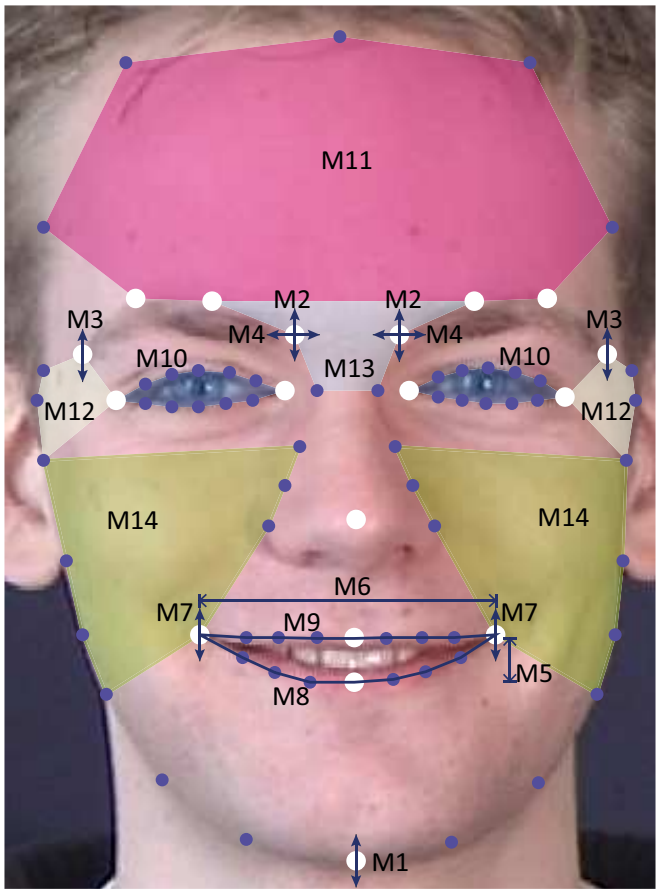


Figure 2: Example of landmarks, geometric features and texture regions. See Table 2 for the details of M1-M14.

3.2. Transformation of the dynamic signals to parameter space

A parametric space is constructed in order to extract dynamic signals from video sequences. When the geometric features, $M1, \ldots, M10$, and texture features, $M11, \ldots, M14$, shown in Table 2 are measured in each frame of the video sequence, the dynamic responses of those features for the subject performing a given expression are obtained. Assume each subject i performs *all* or a *subset* of the six expressions $e \in \{smile, surprise, sadness, anger, disgust, fear\}$ recorded as a video $v_{i,e}$, there are a set of measurements $m_{i,e,j}(t), 1 \leq j \leq 14$, where t indicates the frame in the sequence. The fourteen measurements for the features in Table 2, associated with a given subject and expression are denoted:

$$v_{i,e}(t) =< m_{i,e,1}(t), \ \ldots, \ m_{i,e,j}(t), \ \ldots, \ m_{i,e,14}(t) > \quad (1)$$

Analysing such a large data space, with time-series of various lengths, is a difficult and challenging task. Therefore, each measurement $m_{i,e,j}(t)$ for all t is converted into a space of 23 real valued parameters. Each parameter (p^k) encodes different aspects of the time-series (e.g., their shape or texture). The first two parameters are **length** (p^1) and **peak** (p^2) of the time-series.

Action Units	Action Descriptions	Linked Geometric Features
AU1	Inner brow raiser	M2
AU2	Outer brow raiser	M3
AU4	Brow lower	M2, M4
AU5	Upper lid raiser	M4, M10
AU7	Lid tightener	M4,M10
AU10	Upper lip raiser	M5
AU12	Lip corner puller	M6
AU15	Lip corner depressor	M7
AU16	Lower lip depressor	M8
AU17	Chin raiser	M1
AU20	Lip stretcher	M9, M11
AU23	Lip tighten	M5, M8, M9
AU24	Lip pressor	M5, M8, M9
AU25	Lips part	M5, M8 M9

Table 1: Generic links between the measurements and FACs

After all $m_{i,e,j}(t)$ are linearly interpolated, they are normalized so that they are of equal length (137 frames, the overall maximum in the dataset). This normalization is necessary in order to compute the remaining parameters. Each video sequence is represented by a vector of length 322 (14×23) :

$$< p^1_{i,e,1}, \ldots, p^{23}_{i,e,1}, \ldots, p^k_{i,e,j}, \ldots, p^1_{i,14}, \ldots, p^{23}_{i,e,14} > \quad (2)$$

Reducing the dimensionality of the parameters and normalising the length of measurement of time series allows the utilisation of machine learning approaches and the visual analysis of the data.

the following is a short description of each of the parameters that are used:

- Simple descriptors (p^1 – p^2): the length and peak of the expression.
- Low-order Moments (p^3 – p^6): the four moments used are average, variance, skewness and kurtosis.
- Principle Component Analysis (PCA) (p^7 – p^{10}): Here the four largest PCA coefficients are used as they typically capture the main patterns in the curves.
- Fourier Coefficients (p^{11} – p^{15}): The four largest DFT coefficients were found to be sufficient to capture the variation in the data.
- Polynomial Fitting (p^{16} – p^{18}): A quadratic polynomial is used to describe each measurement. Here three polynomial coefficients are used.

Measurements Index	Measurements Descriptions
M1	Chin vertical disp.
M2	Inner brows vert. disp.
M3	Outer brows vert. disp.
M4	Brows horizontal disp.
M5	Mouth height
M6	Mouth width
M7	Mouth corner vert. disp.
M8	Lower lip curvature
M9	Upper lip curvature
M10	Eye region size
M11	Forehead Gabor response
M12	Eye corner Gabor response
M13	Inner brows Gabor response
M14	Cheeks Gabor response

Table 2: Extracted measurements

- Auto-regressive (AR) Model (p^{19} – p^{23}): The final five parameters are obtained from a least squares AR model.

3.3. Machine learning techniques

For the classification of the facial expressions using the data generated from the process in Section 3.2, a number of different testing, validation and training schemes are employed. The first of these involves manually dividing the data into a training set and an independent test set using a 50% stratified split. This means that the training and test sets have the same proportion of expressions (as far as possible), and number of 'difficult' expressions as the complete dataset. This split results in a training set of 102 objects, and a test set of 101 objects. The second approach involves the use of stratified 10×10-fold cross validation to generate models using all of the data.

The classifier learners employed for this work are drawn from different areas of machine learning, and six such classifiers are used. The reason that such diverse range of classifiers is employed is that the best result can be leveraged from the data and that the results presented are realistic. The six learners utilised for this study are J48 (a version of ID3) [35], FRNN (a Fuzzy-Rough based Nearest Neighbour algorithm) [36], VQNN (Vaguely Quantified Nearest Neighbour, a noise tolerant fuzzy-rough classifier) [6], Random Forest (a tree-based classifier) [7], SMO-SVM (Sequential Minimal Optimisation approach for Support Vector Machines) [37], and Logistic (a multinomial logistic regression classifier) [38] which are described briefly below.

J48 is based on ID3 [35] and creates decision trees by choosing the most informative features and recursively partitioning the data into subtables based on the values of such features. Each node in the tree represents a feature with branches from a node representing the alternative

values this feature can take according to the current subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created to represent this classification.

FRNN [36] is a nearest-neighbour classifier based on fuzzy-rough sets. It uses the fuzzy upper and lower approximation memberships of the test object to its nearest neighbours in order to predict the decision class of a test object. It should be noted that FRNN does not require the specification of the k nearest-neighbours and all neighbours are used in the evaluation.

VQNN [6] is based on vaguely quantified rough sets. This is an approach which uses vague quantifiers to minimise the dominance of noisy features on classification. The approach uses more flexible definitions of the traditional fuzzy upper and lower approximations, thus reducing the influence of extreme-valued features.

Random forest [7] is an ensemble classifier that consists of many randomly-built decision trees. It outputs the decision class for a test object that is the mode of the classes obtained by individual trees.

SMO-SVM [37] is an algorithm for efficiently solving the optimisation problem which arises during the training of support vector machines. It breaks optimisation problems into a series of smallest possible sub-problems, which are then resolved analytically.

Logistic [38] is a classifier that builds a logistic regression model using a multinomial ridge estimator.

4. Experimental evaluation

The evaluation of any expression recognition system is a non-trivial task for a number of different reasons. Firstly, the changes in facial expression are diverse, as they are controlled by complex human emotions and personally distinctive characteristics. Therefore, some expressions are more difficult to distinguish from others. Secondly, the data contained in the publicly available databases are often collected artificially, where subjects have been instructed to mimic expressions. Such artificial mimicry does not contain the associated emotions and this can hinder the objective evaluation of any system. Thirdly, there are a number of parameters attached to the different stages of data generation and classification.

In the work in this paper, three methods are used in order to perform a comprehensive and robust performance evaluation. First, comparison by recognition rate (overall classification accuracy) is a standard way to evaluate performance. This is a general indicator of the efficiency of the system. Secondly, visualisation techniques, which have the ability to visually analyse the outliers in the dataset, are also exploited to evaluate the extracted facial features for facilitating the investigation into misclassified sequences. Thirdly, a user study is carried out in order to investigate those characteristics which are common to both human perception and automatic machine vision systems.

4.1. Data

Even a cursory examination of the literature will show that many different datasets (and indeed subsets of datasets) have been used in previous work for the task of automatic facial expression analysis, [10, 39]. This can make the comparison of various techniques difficult as there is no common frame of reference in which to compare the performance of different methods. In this work, the aim is to make the task of comparison with other methods much easier, and therefore the widely used MMI database [40, 41] is utilised. This dataset is publicly available and has been used for several other publications e.g. [42, 43, 44].

Expression	No. of seqs.
Anger	32
Disgust	28
Fear	28
Happiness	42
Sadness	32
Surprise	41

Table 3: Expression data

The video sequences chosen for inclusion in the data were based on the attached label for the six basic expressions. This resulted in a collection of 203 sequences.[1] As mentioned previously, the aim of this study was to use as much data as possible without any subjective removal of sequences that were considered 'undesirable' (thus making the task of discerning different expressions easier). Having gathered the data it became apparent that not all of the sequences were suitable, as several contained only profile-views of the subject (i.e. do not show the whole face). Having discarded this unusable data, a total of 203 frontal view sequences (together with their associated labels) remained. These expressions are summarised in Table 3.

Further analysis of the 203 video sequences revealed that in some particular examples there is occlusion of the subject or their face, or that there is no visible change in expression throughout the sequence. It was decided *not to remove* these sequences but rather to treat them in the same way as those that were 'good' examples of their relevant assigned label. A number of strategies for dealing with such sequences are presented in a later section including use of human reasoning in the form of a user study. It is important to note that these strategies ensure that video sequences were not removed subjectively or discarded simply because they were difficult to classify.

4.2. Static Recognition vs. Dynamic Recognition

In this section, three types of static recognition methods are used to compare the effectiveness of the proposed dynamic features approach. A RBF kernel SVM-based classifier is selected for the learning step. The first benchmark is Local Binary Patterns and SVM, as used in [22]. The face patch extracted from the frame with peak expression is divided into 6×7 regions for extracting LBP features. The second and third benchmarks are Active Shape Model (ASM) features [45] and Active Appearance Model (AAM) features [28], where we abuse the terms as we only use the feature representations and ignore the search component. These two types of feature representations have been widely used in facial expression recognition and synthesis, e.g. [46].

The same independent training and test scheme was used as that described previously in section3.3.

The results of the overall and individual classification accuracy on the independent test set are shown in Table 4. When investigating the features of static recognition, it is found that shape features, e.g., ASM, has more distinguishing power than texture features, e.g., LBP[22]. Furthermore, the table shows that the proposed dynamic feature outperforms all the other features

[1]The database was accessed at `http://www.mmifacedb.com/` in March 2011. Video sequences were obtained using the form search option on the website and requesting all video sequences which are labelled as either: *anger, disgust, fear, happiness, sadness* or *surprise*.

Features	Overall	anger	surprise	sadness	happiness	fear	disgust
LBP	**54.45**	50.00	35.71	38.46	71.43	62.50	61.90
AAM	**62.38**	62.50	42.86	38.46	80.95	56.25	76.19
ASM	**64.35**	68.75	64.29	46.15	85.71	75.00	42.86
Proposed	**71.56**	75.00	85.00	50.00	90.50	50.00	64.30

Table 4: Comparison of static recognition systems and proposed system with manually stratified training and test data (SMO-SVM [37] is used as classifier)

Classifier	Overall	anger	surprise	sadness	happiness	fear	disgust
J48	**50.00**	68.80	76.20	18.80	52.40	35.70	35.70
FRNN	**71.57**	43.80	90.50	75.00	95.20	57.10	50.00
VQNN	**70.58**	43.80	81.00	87.50	90.50	57.10	50.00
RF	**57.84**	56.30	76.20	43.80	85.70	35.70	28.60
SMO-SVM	**71.56**	75.00	85.00	50.00	90.50	50.00	64.30
Logistic	**69.60**	81.30	85.70	43.80	85.70	42.90	64.30

Table 5: Classification accuracy (%) with manually stratified training and test data

extracted from static image with peak expression to achieve 71.56 % which is the highest recognition accuracy rate. The result indicates that dynamic features extracted from sequences are more suitable for the task of facial expression recognition.

4.3. Classifier learning

A number of experiments were carried out using the dataset obtained by extracting the dynamic signal data from the 203 video sequences as described in Section 3.2. This evaluation is divided into three parts. The first examines the data after it has been manually stratified and divided into independent testing and training sets. The second part uses 10×10-fold cross validation to generate models. The third examines those sequences that are consistently misclassified by all of the classifier learners in the first part, and tries to reason about the results.

4.3.1. Classifier learning with manually stratified training and testing data

Generally there are two opposing views regarding the use of cross validation for model selection and validation [47]. One view holds that an independent test set must *always* be used in order to ensure that there are no *a-priori* similarities between those objects in the training data and those of the test data [47]. However, the examination of the data in order to divide it into test and training sets is in itself a violation of that independence, and this forms the basis for the opposing view [48]. In order to avoid such pitfalls, in this work, each of these training/testing schemes have been implemented and results are presented for both.

The results for the overall classification accuracy for the independent test set, and the accuracy for each class using the six previously described classifier learners is shown in Table 5. In order to generate a robust classification model from the testing data, the training set was first validated using a 10-fold cross validation. This means that the training data was trained and validated only on the training data by doing an internal validation. The resulting averaged prediction model was then used to classify the independent test set data.

Classifier	Overall (SD)	anger	surprise	sadness	happiness	fear	disgust
J48	**51.92 (9.06)**	50.00	65.90	46.90	68.90	35.70	30.00
FRNN	**75.96 (9.22)**	65.60	82.90	75.00	91.10	67.90	63.33
VQNN	**69.71 (8.90)**	43.80	82.90	81.30	97.80	35.70	56.70
RF	**62.98 (10.76)**	65.60	82.90	53.10	91.10	21.40	40.00
SMO-SVM	**70.67 (9.70)**	68.80	78.00	56.30	93.30	42.90	70.00
Logistic	**50.96 (9.03)**	50.00	51.20	46.90	64.40	28.60	56.70

Table 6: Classification accuracy (%) with stratified cross-validation

It is clear from the results shown in table 5 that FRNN, VQNN and SMO-SVM offer the best overall performance. What is also apparent is that amongst all classifiers *happiness* and *surprise* appear to be the easiest expressions to classify. Although, J48 does have difficulty in achieving the same performance as other learners for *happiness*. This performance is easy to explain since both of these expressions are the best represented in the dataset with 42 and 41 data objects respectively. The expression *anger* seems to be difficult for most learners, but the Logistic approach and SMO-SVM do well here with accuracies of 81.3% and 75% respectively. J48 also manages to return almost 69%. The expression *disgust* also appears to offer rather mixed results with SMO-SVM and logistic returning results of around 64.3% while FRNN and VQNN do less well with around 50%. Note that the value for k used for VQNN was 7, and no attempt was made to 'tune' this. It is possible that other values for k would result in better performance.

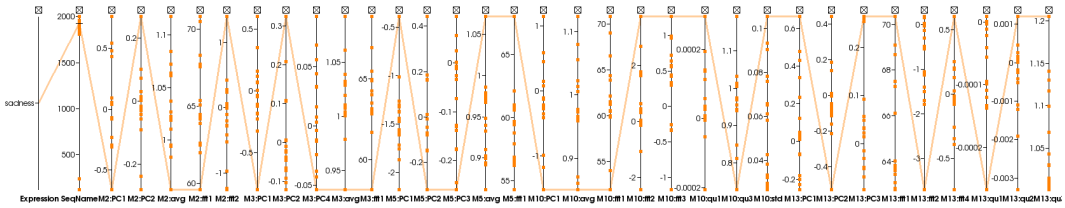
For the test data, 26 of the 102 objects are consistently predicted correctly by *all* six classifiers, whilst eight sequences are consistently misclassified by all six classifiers. In order to provide an assessment of the misclassified video sequences, each of those sequences were examined individually and were then employed for the user study, documented in section 4.5.

4.3.2. Classifier Learning with stratified cross-validation

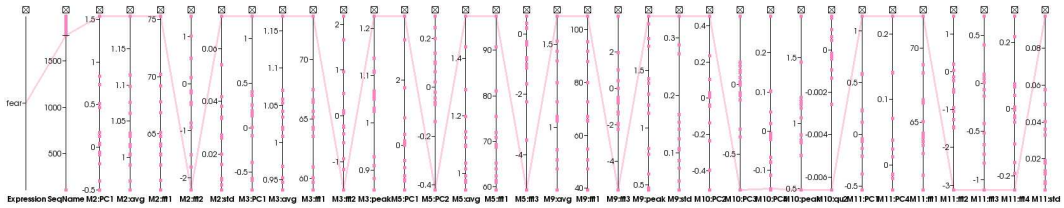
This set of experiments were conducted using *all* of the data, and testing/training is performed as part of the internal 10-fold cross-validation. This was carried out in order to gain an understanding of how stable classifiers could be obtained and what results could be expected. Using a 10 times 10 fold cross-validation approach, a number of experiments were carried out. In Table 6 it can be seen that there is an increase in classification for some approaches, but as the testing and training phases are intertwined it is more difficult to extract those misclassified sequences. What is clear however is that 10 fold cross-validation reduces the tendency for extreme values and in most classifiers strengthens those results which are poor for the manually stratified data in the previous section. Once again it is apparent that *fear* and *disgust* are the most difficult expressions to classify. What is more interesting is that this scheme allows the examination of the variation of the model stability for different randomisations of the data, expressed as sd values over the 10 runs. It can be seen that FRNN and VQNN do particularly well in these cases when overall classification accuracy is taken into account.

4.4. Misclassified sequence analysis

During the previous evaluation it was discovered that some video sequences are consistently mis-classified by all classifiers. Since there are more than 300 facial feature measurements, it can be difficult to manually ascertain the reasons for this misclassification. In an attempt to visualise this aspect, a tool [49] which uses parallel coordinates and scatterplots to analyze these



(a) 1926 Sadness



(b) 1770 Fear

Figure 3: Example visualizations of facial feature measurements.

measurements is employed. Parallel coordinates visualisation is known to be useful for identifying clusters, separation and outliers in high dimensional data space. The use of scatter plots in conjunction with parallel coordinates has been shown to be useful for cluster identification [50].

For the work described here, the tool is adapted such that the first parallel coordinate axis shows the classification of the sequence under consideration. The remaining axes represent the values of these measurements. The scatterplots (dot points on each axis) show the distributions of measurements of the whole class. All of the measurements which belong to the highlighted sequence are connected with a line across all the parallel coordinate axes. This provides an intuitive understanding of how the measurements of the highlighted sequence are distributed relative to other sequences in the same group. Two examples are shown in Figure 3.

An important observation is that many facial feature measurements of these mis-classified sequences lie on the boundaries relative to those of its labelled class. For example, in Figure 3, a sadness sequence (id: 1926) and a fear sequence (id: 1770) are found to have a significant number of measurements bordering the tip / bottom of the axes. Each of the axes are normalized using the maximum and minimum values. Due to space constraints, only a limited number of visualisations are shown. Such a large number of extreme measurements suggest that there may be problem extracting features for these sequences or that these sequences may even be incorrectly labelled [49]. In order to further investigate these sequences, this aspect is examined as part of the user study to determine whether humans can classify these sequences correctly.

4.5. User study

In order to produce a human evaluation of the machine learning methods, a user study is devised which allows comparison with human reasoning. The evaluation is focused on those particular stimuli which are consistently misclassified by all six machine learning methods. A total of 16 video sequences are selected: eight which are consistently misclassified and eight which are consistently correctly classified. To increase stimuli reliability the sequences chosen that are correctly classified underwent scrutiny by a small but ethnically diverse group of people.

4.5.1. Participants.

A total of 11 participants (4 female, 7 male) took part in this experiment in return for a GBP£10 book voucher. Participants belonged to the *Swansea University* student community and with a diverse variety of disciplines including Humanities, Engineering and Economics. Ages ranged from 18 to 22 (mean=21.7, sd=0.8). All participants had normal or corrected to normal vision and were not informed about the purpose of the study at the beginning of the session.

4.5.2. Apparatus.

Stimuli consisted of video sequences from the MMI database [40] and were presented to participants using a custom made interface. Experiments were run using Intel Dual-Core PCs running at 2.13 GHz, with 2 GB of RAM and Windows 7 Professional. The display was 19in LCD at 1280x1024 resolution with a 32bit sRGB colour mode. Each monitor was adjusted to have same brightness and same level of contrasts. Participants interacted with the software using a standard mouse at a desk in a dimmed experimental room.

4.5.3. Task and procedure.

The experiment began with a brief overview read by the experimenter using a predefined script. Detailed instructions were then given through a self-paced slide presentation. Brief descriptions of the requirements of the task were also provided. The experiment consisted of a single task in which each participant was asked to classify a video sequence according to one of the 6 classes provided: anger, surprise, sadness, happiness, fear, disgust. Specific instructions were given onscreen before each video sequence was shown and a total of 6 practice trials were also completed to familiarise participants with the interface. A blank screen was shown for 10 seconds before each stimulus was presented to refresh participants short term memory. At the end of each trial the task would enter a holding state waiting for the participant to press a NEXT button (whenever he/she felt comfortable) which would allow the experiment to proceed to the evaluation of the next stimulus. When the experiment had been completed each participant completed a short debriefing questionnaire and was provided with information about our experimental goals.

4.5.4. Results and discussion

The results are shown in table 7, where 1 denotes agreement with the label originally assigned to that particular sequence in the dataset. Conversely, 0 indicates disagreement. The agreement/disagreement rate with the assigned labels for the classifier learners is also provided as well for the 11 participants. Note that the data sequences used for this study included eight sequences where all six classifier learners consistently *disagreed* with the assigned label and eight where they all consistently *agreed* with the given label. The last line of the table provides a summary of agreement/disagreement for all participants. Based on the analysis of this user study, it was concluded that expression recognition based on automatic learning methods is highly correlated with human perception. This is reflected in the following observations:

- The mean agreement between the (consistently correct) automatically learned labels and the human participants is 92.04% (sd = 3.63). This indicates that those sequences that were always correctly predicted by the automatic methods also had an average of 92% of agreement amongst all of the human participants.

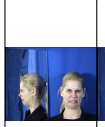
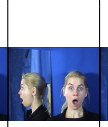
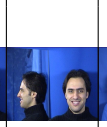
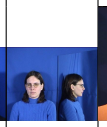
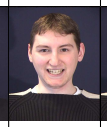
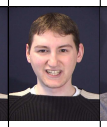
Assigned label	disg	sup	ang	hap	ang	fear	sur	disg	sad	disg	fear	sad	fear	disg	disg	hap
Video ID	104	117	218	225	686	1770	1771	1822	1863	1882	1922	1926	1962	1965	1992	1996
																
ML	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1
P1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1
P2	1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	1
P3	1	1	0	1	0	0	1	1	1	1	0	1	0	0	1	1
P4	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1
P5	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1
P6	1	1	1	1	0	0	1	1	1	1	0	1	0	0	0	1
P7	1	1	1	1	0	0	1	1	1	0	0	0	1	0	0	1
P8	1	1	1	1	0	1	0	1	1	1	0	0	0	0	0	1
P9	0	1	0	1	0	0	1	1	1	1	1	0	0	1	1	1
P10	1	1	0	1	0	0	1	1	0	1	0	1	1	0	1	1
P11	1	1	0	1	0	0	1	1	1	1	1	1	0	0	0	1
Summ.	**10/1**	**11/0**	**7/4**	**11/0**	**10/1**	**10/1**	**10/1**	**11/0**	**10/1**	2/9	**7/4**	4/7	**8/3**	**9/2**	**7/4**	**11/0**

Table 7: User study: comparison of automatic learning with human perception. 'ML' refers to the labels automatically learnt by the machine learning algorithms. Rows P1-P11 indicate whether human participants agreed or disagreed with the given labels. 'Summary' details the number of agreements/disagreements between the automatic methods and human labels. (bold).

- The mean agreement between those sequences that are consistently incorrectly classified by the automatic methods and the human participants is 64.76% (sd = 68.55). This reflects the confusion amongst the different automatic methods for these particular sequences that is also experienced in human observers. In particular, this is borne out by the high sd values indicating a distribution with large extremes and hence poor agreement.

- If a simple majority vote is used to summarise the results of the human reasoning, it is shown that human participants agreed with the automatic classifications in 14 sequences out of the 16 (total) sequences.

- When the sequences with the most discrepant answers were selected from table 7, they include 2 *fear*, 1 *anger*, 1 *disgust* and 1 *sadness*. In tables 5 and 6, it is demonstrated that most of the automatic learning methods achieve consistently high accuracy for the expressions *surprise* and *happiness* whilst there was significant variation for the other expressions (it should be noted however that the *surprise* and *happiness* expressions are the most well represented in the dataset considered in this paper). This relative ease of classification is also reflected in this user study, as the human participants were able to identify the *surprise* and *happiness* expressions easily but relied heavily on contextual information (e.g., body movements) to classify the other four.

5. Conclusion

Facial expression analysis and recognition has become one of the most active research topics in recent decades due to its potential contribution to future human-computer interaction analysis. In this paper, a data-driven approach was proposed in order to exploit the dynamic information in video sequences for automatic expression recognition. This was achieved by generating a facial landmark tracking framework and building a parametric space in which to capture the dynamic signal information from both the geometric and texture features. A comprehensive range of machine learning methods were then employed for the task of facial expression recognition. A robust approach such as this to the evaluation step not been presented previously in the literature.

The evaluation aspect of the work was further developed by including a framework for classification accuracy comparison, feature visualisation, and also by offering a novel correlation analysis of human perception and machine vision through the use of a user study. This multi-faceted evaluation provides an intuitive way to guide future work on facial expression analysis and in particular recognition. In the evaluation, both automatic classifiers and human participants were able to classify the expressions of *happiness* and *surprise* easily, but encountered difficulty in identifying the other basic expressions. However, the dataset used in this paper is relatively small, and some of the expressions are poorly represented (e.g. *disgust* and *fear*), whilst others are well represented. This will make it difficult for classifier learners to learn a given concept well. One way to address this is to either balance the dataset, or acquire more data. Other aspects that could be useful perhaps are the use of more contextual information such as audio data and body movements could in order to achieve better performance and a better understanding of human emotions.

Topics for future work include: the investigation of the correlation between the automatic learning methods, the integration of contextual information for expression recognition, and the investigation of the applicability of the work to other forms of video media (skype, video conferencing, streamed data, etc.).

References

[1] C. Darwin, The Expression of the Emotions in Man and Animals, J. Murray, London, 1872.

[2] P. Ekman, W. Friesen, Constants across cultures in the face and emotion, Journal of Personalisty Social Psychology 17 (1971) 124–129.

[3] N. Suwa, M. Sugie, K. Fujimora, A preliminary note on pattern recognition of human emotional expression, Proc. Int. Joint Conf. on Pat. Rec. (1978) 408–410.

[4] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comp. Vis. 57 (2004) 137–154.

[5] H. Fang, N. P. Costen, From rank-n to rank-1 face recognition based on motion similarity, in: Proc. British Conf. on Mach. Vis., pp. 1–11.

[6] C. Cornelis, M. D. Cock, A. M. Radzikowska, Vaguely quantified rough sets, in: Proc. Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, volume 4482, Springer, 2007, pp. 87–94.

[7] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[8] M. Pantic, I. Patras, Detecting facial actions and their temporal segments in nearly frontal-view face image sequences, in: Proc. IEEE Int. Conf. Systems, Man and Cybernetics, volume 4, pp. 3358 – 3363 Vol. 4.

[9] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: Machine learning and application to spontaneous behavior, in: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., IEEE Computer Society, Washington, DC, USA, 2005, pp. 568–573.

[10] Y.-I. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pat. Ana. & Mach. Int. 23 (2001) 97–115.

[11] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Trans. Pat. Ana. & Mach. Int. 22 (2000) 1424–1445.

[12] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (2003) 259–275.

[13] Q. Zhang, Z. Liu, B. Guo, D. Terzopoulos, H.-Y. Shum, Geometry-driven photorealistic facial expression synthesis, IEEE Trans. Vis. & Comp. Graphics 12 (2006) 48–60.

[14] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: Audio, visual and spontaneous expressions, IEEE Trans. Pat. Ana. & Mach. Int. 31 (2009) 39–58.

[15] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pat. Ana. & Mach. Int. 27 (2005) 699–714.

[16] M. Pantic, I. Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Systems, Man, and Cybernetics, Part B 36 (2006) 433–449.

[17] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pat. Ana. & Mach. Int. 29 (2007) 1683–1699.

[18] Y. Zhang, Q. Ji, Z. Zhu, B. Yi, Dynamic facial expression analysis and synthesis with mpeg-4 facial animation parameters, IEEE Trans. Circuits and Systems for Video Technology 18 (2008) 1–15.

[19] G. Donato, M. Bartlett, J. Hager, P. Ekman, T. Sejnowski, Classifying facial actions, IEEE Trans. Pat. Ana. & Mach. Int. 21 (1999) 974–989.

[20] M. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Systems, Man, and Cybernetics, Part B 42 (2012) 28 –43.

[21] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proc. IEEE Conf. Automatic Face and Gesture Recognition, pp. 46 –53.

[22] C. Shan, S. Gong, P. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image and Vision Computing 27 (2009) 803–816.

[23] P. Ekman, E. L. Rosenberg (Eds.), What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (FACS), Oxford University Press, Oxford New York, 2005.

[24] I. Kotsia, I. Pitas, Facial expression recognition in image sequences using geometric deformation features and support vector machines, IEEE Trans. Image Processing 16 (2007) 172–187.

[25] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pat. Ana. & Mach. Int. 29 (2007) 915–928.

[26] I. Cohen, N. Sebe, A. Garg, L. S. Chen, T. S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vis. Image Underst. 91 (2003) 160–187.

[27] M. F. Valstar, H. Gunes, M. Pantic, How to distinguish posed from spontaneous smiles using geometric features, in: Proc. Int. Conf. on Multimodal interfaces, pp. 38–45.

[28] T. Cootes, G. Edward, C. Taylor, Active appearance models, IEEE Trans. Pat. Ana. & Mach. Int. 23 (2001) 681–685.

[29] D. Cristinacce, T. Cootes, Automatic feature localisation with constrained local models, Pattern Recognition 41 (2008) 3054–3067.

[30] P. Tresaderrn, H. Bhaskar, S. Adeshina, C. Taylor, T. Cootes, Combining local and global shape models for deformable object matching, in: Proc. British Conf. on Mach. Vis., pp. 1–12.

[31] T. Cootes, C. Twining, V. Petrovic, K. Babalola, C. Taylor, Computing accurate correspondences across groups of images, IEEE Trans. Pat. Ana. & Mach. Int. 32 (2010) 1994–2005.
[32] A. Aubrey, V. Kajicc, I. Cingovska, P. Rosin, D. Marshall, Mapping and manipulating facial dynamics, in: Proc. IEEE Conf. Automatic Face and Gesture Recognition, pp. 639–645.
[33] M. Bartlett, P. Viola, T. Sejnowski, J. Larsen, J. Hager, P. Ekman, Classifying facial action, in: Proc. Neural Information Processing Systems, volume 8, pp. 823–829.
[34] M. Pantic, L. J. M. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Trans. Systems, Man, and Cybernetics, Part B 34 (2004) 1449–1461.
[35] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[36] R. Jensen, C. Cornelis, A new approach to fuzzy-rough nearest neighbour classification, in: Proc. Int. Conf. on Rough Sets and Current Trends in Computing, volume 5306, Springer, 2008, pp. 310–319.
[37] J. C. Platt, Fast training of support vector machines using sequential minimal optimization, in: Adv. in kernel methods, MIT Press, Cambridge, MA, USA, 1999, pp. 185–208.
[38] S. Le Cessie, J. C. Van Houwelingen, Ridge estimators in logistic regression, Appl. Stat. 41 (1992) 191–201.
[39] C. Juanjuan, Z. Zheng, S. Han, Z. Gang, Facial expression recognition based on PCA reconstruction, in: Proc. Int. Conf. Computer Science and Education, pp. 195–198.
[40] M. Pantic, M. F. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proc. IEEE Int. Conf. Multimedia and Expo, Amsterdam, The Netherlands, pp. 317–321.
[41] M. F. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proc. Int. Conf. Language Resources and Evaluation, Workshop on EMOTION, pp. 65–70.
[42] R. Contreras, O. Starostenko, V. Alarcon-Aquino, L. Flores-Pulido, Facial feature model for emotion recognition using fuzzy reasoning, in: Proc. Conf. on Pat. Rec.: Adv. in Pat. Rec., Springer-Verlag, Berlin, Heidelberg, 2010, pp. 11–21.
[43] S. ur Rehman, L. Liu, H. Li, Lipless tracking and emotion estimation, in: Proc. Int. Conf. Signal-Image Technologies and Internet-Based System, pp. 768–774.
[44] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Trans. Pat. Ana. & Mach. Int. 32 (2010) 1940–1954.
[45] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models - their training and application, Comput. Vis. Image Underst. 61 (1995) 38–59.
[46] B. Abboud, F. Davoine, M. Dang, Facial expression recognition and synthesis based on an appearance model, Signal processing: image communication 19 (2004) 723–740.
[47] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, Journal of Chemometrics 24 (2010) 194–201.
[48] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Morgan Kaufmann, 1995, pp. 1137–1143.
[49] G. Tam, H. Fang, A. Aubrey, P. Grant, P. Rosin, D. Marshall, M. Chen, Visualization of time-series data in parameter space for understanding facial dynamics, Computer Graphics Forum 30 (2011) 901–910.
[50] D. Holten, J. J. Van Wijk, Evaluation of cluster identification performance for different PCP variants, Computer Graphics Forum 29 (2010) 793–802.