Is the 2D Unlabelled Data Adequate for Facial Expression Recognition?

Lili Tao[a,b] and Bogdan J. Matuszewski[a]

***Abstract* — Automatic facial expression recognition is one of the important challenges for computer vision and machine learning. Despite the fact that many successes have been achieved in the recent years, several important but unresolved problems still remain. This paper describes a facial expression recognition system based on the random forest technique. Contrary to the many previous methods, the proposed system uses only very simple landmark features, with the view of a possible real-time implementation on low-cost portable devices. Both supervised and unsupervised variants of the method are presented. However, the main objective of the paper is to provide some quantitative experimental evidence behind more fundamental questions in facial articulation analysis, namely the relative significance of 3D information as oppose to 2D data only and importance of the labelled training data in the supervised learning as opposed to the unsupervised learning. The comprehensive experiments are performed on the BU-3DFE facial expression database. These experiments not only show the effectiveness of the described methods but also demonstrate that the common assumptions about facial expression recognition are debatable.**

***Index Terms*— Facial expression recognition, random forest, non-linear manifold learning, supervise and non-supervised learning.**

1. Introduction

FACIAL expression analysis has attracted a significant research interest during past several years due to its importance for providing cues helping to understand human behaviour, analyse emotions and assess intentions. As an active research field with extensive applications in many different areas, large body of literature exists on 2D/3D static and dynamic recognition systems, with significant progress made towards achieving high recognition rate. De la Tore et al. [1] provide a comprehensive overview of methods summarising the fundamental approaches and the recent advances in automatic facial expression analysis from 2D intensity images or video sequences. Sandbach et al. [2] offer a survey describing the use of both static and dynamic 3D data. Facial expression recognition systems are typically composed of two subsystems: feature extraction and feature classifier. Many publications mainly focus on extracting sophisticated highly discriminative facial features. These features can be either hand-designed or learned from the training data. It is known that some features are more critical for analysing facial expressions than the others and the feature selection procedure can be applied to improve the performance [3], [4]. Indeed, extracting complex 2D or 3D features can improve the systems performance, but often requires more computational resources. This may not be acceptable for real-time applications particularly run on inexpensive portable devices. It is also often claimed, when 3D information is being used, that due to the lack of depth information, 2D data is not suitable to represent intrinsic facial structure and therefore not proper for complex facial expressions recognition. However, 3D information is still expensive to collect and is not available for many scenarios. With very few exceptions, most reported facial expression recognition systems are based on supervised learning, which requires labelled data in the training process. Very little attention has been paid to the unsupervised systems. The work in [5] clusters the similar facial events using an unsupervised learning, but only works for small number of subjects. Considering the time and cost involved in producing the labelled data as well as often questionable quality of such ground truth, an unsupervised system would be particularly useful. For facial expression recognition the popular classification algorithms, include: Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM), often combined with a boosting algorithm for feature selection [6]. The discriminant function in LDA has an intuitive interpretation as it maximises between-class and minimises within-class scatter, but only handle the data when the relation between them is linear. Although SVM is very successful, it is intrinsically designed to solve binary classification problems. Although it has been adapted to work with multiple classes, one-vs-all the SVM approach may lead to asymmetries which are not really justified by the training data [7]. On the other hand, a random forest is naturally designed for solving multi-class classification problems with an additional uncertainty encoded in its probabilistic output. Such techniques have become very popular recently given their capability to provide good discrimination, to reduce over-fitting, and enabling simple parallel implementation. In this work, simple landmark features are used for the facial expression recognition in both supervised and unsupervised approaches. Recently a number of methods have been proposed for efficient, robust and accurate 2D facial landmark detection and tracking [8] including commercial products [9], making them feasible for application on portable devices.

Although similar questions on the significance of the 3D versus 2D data have been addressed in some papers, these are based on rather limited tests. Furthermore, the majority of these

a. Robotics and Computer Vision Research Laboratory, School of Computing, Engineering and Physical Sciences, University of Central Lancashire, UK
b. Visual Information Laboratory, Department of Computer Science, University of Bristol, UK

papers deal with the face recognition. The authors have been unable to identify published papers addressing these issues in a systematic way for the facial expression recognition problems. For example, in the recently published paper [10] a limited comparison between the 3D and 2D datasets has been performed as part of a validation process of the developed comprehensive database, but with 2D images showing only the frontal faces. In [11] a simple comparison has been made between 3D technique and 2D based methods with the 2D data rendered directly from the 3D data – this though have been done without error analysis, e.g. due to environmental changes. Additionally only the labeled data was used in these tests.

Instead of developing the "best method" that outperforms the state-of-the-art, the purpose of this paper is mainly to explore more fundamental but important questions which have rarely been investigated in the previous works. This is not to say that all the possible experimental configurations have been investigated or all questions answered, indeed far from this, but the authors hope that the reported results and the proposed methodology are of general interest and are robust enough as to assist in the further discussions of these topics.

The paper is structured as follows: Section 2 highlights the contributions of the paper. Section 3 presents the methods used for facial expression recognition, justifying its selection. An extensive set of experiments is described in Section 4 to address two main questions: (i) are the 3D landmarks, as opposed to 2D landmarks only, significant for improving facial expression recognition, (ii) is the availability of the labelled training data really significant. This section also includes a link to some previously published results. The discussions of the results are provided in Section 5.

2. CONTRIBUTIONS

This paper presents a random forest based subject-independent facial expression recognition system for six prototypical emotions: anger, disgust, fear, happiness, sadness and surprise, using both supervised and unsupervised approaches. In the supervised method random decision forest is used to perform the multiclass classification. In the unsupervised setting, the density forest is employed to identify the local neighbourhood structures in the feature space subsequently used to calculate the affinity matrix defining diffusion maps manifold. Contrary to the most existing papers, instead of putting effort on extracting and selecting complex features, the focus of this paper is on using simple landmark features, and assessing how well the proposed recognition system can deal with the problem in such case.

The important contribution of this paper is the exploration of more fundamental questions: whether, in case of used simple features, 3D information is significantly helpful for recognising specific facial expressions. Is the labelled training data really needed or is it possible to build an unsupervised system having comparable performance to the supervised facial expression recognition system? Given 2D data only, is maintaining a consistent facial pose necessary for achieving good recognition performance? To the best of authors' knowledge, the quantitative consideration of such questions is rarely provided in the previous works.

3. METHODOLOGY

To answer the questions presented in Section 2, this section describes the proposed methods used for facial expression recognition. The simple landmark features are described first, followed by the details of the random forest classifier. The proposed implementation is based on the random forest classification and manifold forest presented in [7]. The use of the proposed random forest methodology with simple facial landmarks features is considered as a good compromise between performance and flexibility of the methodology enabling consistent tests for different considered scenarios leading to robust and compact results which could be reported in a short paper.

3.1 Feature description

Given a set of face features $\mathcal{F} = \{\mathcal{F}^1 \quad \dots \quad \mathcal{F}^N\}$ representing *N* different subjects with each subject having *F* faces in the database and each face described by features derived from *P* landmarks. In this paper, 83 landmarks are used as defined in the BU-3DFE database [12]. Each subject, is represented by the feature set $\mathcal{F}^k = \{\mathbf{F}_1^k \quad \dots \quad \mathbf{F}_F^k\}$, where $\mathbf{F}_j^k \in \mathbb{R}^{D\times 1}$ is the feature vector representing face *j* of subject *k*, and *D* is the dimension of the feature vector. The feature vector is defined as the difference between all the landmarks' position of the given face and the corresponding landmarks' position of the reference face showing subjects' natural expression: $\mathbf{F}_j^k = \mathbf{S}_j^k - \mathbf{S}^k$, where $\mathbf{S}^k$ is the neutral expression face vector of subject *k*. Each face is represented by a face vector: $\mathbf{S}_j^k = [\mathbf{p}_{1j}^k \quad \dots \quad \mathbf{p}_{Pj}^k]^T \in \mathbb{R}^{D\times 1}$, where $\mathbf{p}_{ij}^k$ is a row vector representing coordinate of i^{th} landmark either in 3D or 2D, thus *D* = *3P* for 3D data and *D* = 2*P* for 2D data.

3.2 Supervised random forest classification

In the supervised system, given a set of extracted features $\mathcal{F}$ from the training data together with training labels $\mathcal{C}$, the objective is to build suitable classifier. In this paper the random decision forest is used as a classifier. In the forest the trees are built by randomly selecting single feature (a randomly selected entry in the feature vector **F**) at each internal node. The data reaching the decision node is assigned to its left or right child

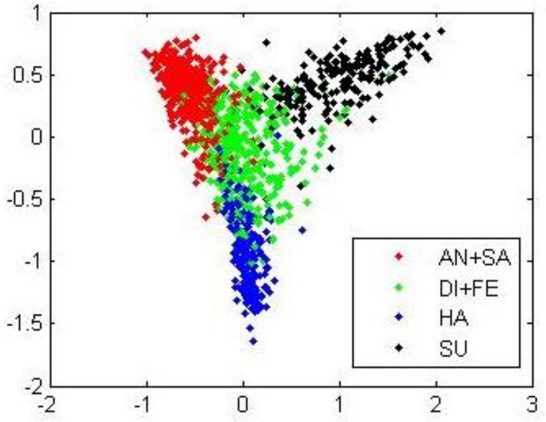


Figure 1. Embedding of the facial expressions data in the 2-dimensional diffusion maps space.

node according to the results of the decision function. The threshold α_m of the decision function at node *m* is selected as a result of the maximisation of the information gain: $\alpha_m^* = \arg\max_{\alpha_m} I_m$, where the information gain is defined as,

$$I_m = H\left(\mathcal{F}_m\right) - \sum_{i\in\{L,R\}} \frac{\left|\mathcal{F}_m^i\right|}{\left|\mathcal{F}_m\right|} H\left(\mathcal{F}_m^i\right) \qquad (1)$$

where $|\cdot|$ indicates a cardinality for the dataset. $\mathcal{F}_m$ denotes the training data $\mathcal{F}$ reaching node *m*, and $\mathcal{F}_m^L$, $\mathcal{F}_m^R$ are the subsets assigned to the left and right child nodes. The entropy is defined as,

$$H\left(\mathcal{F}_m\right) = -\sum_{c\in C} p(c)\log(p(c)) \qquad (2)$$

where $\mathcal{C}$ represents a set of all classes, and *p*(*c*) is the probability of class *c*.

After forest is trained, a new sample can be simply put through each tree. Depending on the result of the decision function at each internal node, the new data is sent to the left or right child node until it arrives at a leaf containing posterior probability of the data belonging to the specific class. The final decision is made based on the average of the responses from all the trees in the forest [7].

3.3 Unsupervised manifold forest clustering

In the unsupervised system, a collection of training data is given in the absence of class labels. It is assumed that data is adequately represented by the Gaussian distributions. In that case the entropy $H\left(\mathcal{F}_m\right)$ in Equation (2) can be calculated analytically as:

$$H\left(\mathcal{F}_m\right) = \frac{1}{2}\log\left((2\pi e)\left|\Lambda(\mathcal{F}_m)\right|\right) \qquad (3)$$

where $\Lambda(\mathcal{F}_m)$ is the covariance matrix of $\mathcal{F}_m$. In this case, the data with relatively high dimensional structure, $\mathcal{F} \in \mathbb{R}^D$, is hard to be represented or analysed, but such complex data might by governed by a small number of parameters. Once the trees have been built, a parameter-free binary affinity model is applied in the proposed method: if two samples end up at the same leaf node of the given tree, the entry of the affinity matrix $\mathbf{W}^t$ for tree *t* is set to 1, and to 0 otherwise. Thus for the ensemble of *T* trees the affinity matrix $\mathbf{W}$ is calculated by averaging over all affinity matrices from each single tree: $\mathbf{W} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{W}^t$.

The manifold forests are constructed upon diffusion maps [13] with the neighbourhood topology learned through random forest data clustering. The diffusion maps technique has the capability to recover underlying structure of a complex manifold, thus is used for mapping the data from a high *D* dimensional space to a reduced, *d* dimensional space, $d \ll D$. The optimal embedding Ψ is defined via eigenvalues λ and its corresponding eigenvectors φ of the Laplace-Beltrami operator [13], such as,

$$\Psi(\mathbf{F}^n) \mapsto \left[\lambda_1\varphi_1(\mathbf{F}^n), \cdots, \lambda_d\varphi_d(\mathbf{F}^n)\right]^T \qquad (4)$$

Once the features have been embedded into the low-dimensional space, a Gaussian mixture model (GMM) algorithm is applied to cluster them into pre-defined number of classes. Figure 1 illustrates the embedding of the training data in the 2-dimensional reduced space.

The embedding function Ψ only provides a mapping for the samples which are included in the given training set. For a new data, its' location in the manifold needs to estimated, an efficient way is to interpolate out-of-sample data onto the learned lower-dimensional feature space, rather than re-training the whole manifold. For each new sample, such interpolation can be calculated based on the Nyström extension [14].

3.4 Missing data

The random forests can be easily adopted to handle the cases with outliers and missing data. Many of more advanced facial landmark detection techniques automatically recognise outliers not returning the corresponding landmarks. Therefore, for brevity of the presentation, only the missing data problem is further investigated in this paper, as the outliers problem can be often reduced to the missing data problem. In the paper it is assumed that the landmarks are only missing in a test set, and so the missing entries could be predicted based on the available training data. In the proposed approach the missing values are replaced by the corresponding training set averages calculated separately for each class , that is, the data with at least one missing entry is replicated *C*-times, where *C* is the number of classes. Subsequently all these modified versions of the data are put through the random forest and the final decision is made based on the average of the responses from the forest for all the amended versions of that data.

4. Experimental results

The performance of the random forests based methods for facial expression recognition is tested on the BU-3DFE database [12]. The database consists of the neutral expressions and 6 basic prototypic expressions each with 4 levels of expression intensity. 90 subjects from the database are used in the experiments, and all the experiments are performed using 9-fold cross-validation scheme. For all the tests the data from the same subject is only used for training or testing, never for both. All the results shown in this section are in percentages.

Table 1 lists the results of using the random forest classifier (RF) against two commonly used classifiers: Support Vector Machine (SVM) (libsvm [15] implementation was used in the experiments) and Linear Discriminant Analysis (LDA) under supervised manner. Based on the results shown in the table, RF provide better overall recognition rate and outperform the other two methods for most facial expressions. *It should be emphasized again that the purpose of the paper is not to propose a new "best" method, but to investigate the effect of using 2D and 3D data*. The random forest methodology is selected as it provides robust results, and is flexible, i.e. it is inherently designed to deal with a multi-class classification problem, is easily adopted to solve clustering problem and effectively handles missing data.

	AN	DI	FE	HA	SA	SU	Overall
SVM	**83.06**	72.22	59.44	82.50	75.28	87.78	76.71
LDA	72.78	71.39	**62.78**	80.83	80.28	87.22	75.88
RF	77.50	**73.06**	53.06	**93.33**	**83.61**	**95.28**	**79.31**

Table 1. Comparison of the proposed random forest classifier with SVM and LDA classifiers.

	AN	DI	FE	HA	SA	SU
AN	77.50	2.50	1.94	2.22	15.83	0.00
DI	8.06	73.06	6.39	5.56	2.50	4.44
FE	2.22	5.83	53.06	18.61	8.89	11.39
HA	1.11	0.56	4.72	93.33	0.28	0.00
SA	13.06	1.67	1.11	0.28	83.61	0.28
SU	0.28	0.56	2.22	0.00	1.67	95.28

	AN+SA	DI+FE	HA	SU
AN+SA	87.22	12.41	0.19	0.19
DI+FE	5.37	84.07	4.44	6.11
HA	0.00	17.78	82.22	0.00
SU	0.00	9.26	0.00	90.74

Table 2. (Top) Confusion matrix for 3D data in supervised learning. The average recognition rates are 79.31%. Total 1000 trees are used in the forest with the maximum the depth 12. (Bottom) Confusion matrix for 3D data in unsupervised learning. The average recognition rates are 85.93%. Total 1000 trees are used in the forest with the maximum tree depth 8.

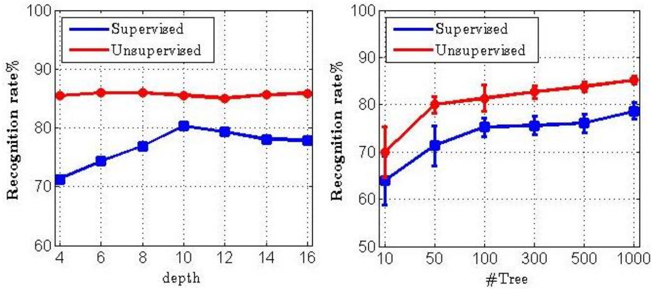


Figure 2. Effect of forest parameters. The average recognition rate (%) as function of: varying tree depth (left), and number of trees in the forest (right).

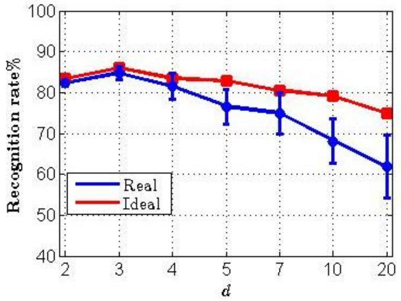


Figure 4. The average recognition rate (%) and the standard deviation as function of dimensionality of the reduced space. The standard deviation was calculated based on 10 experiments with random initialisation of the GMM algorithm.

4.1 Forest parameters

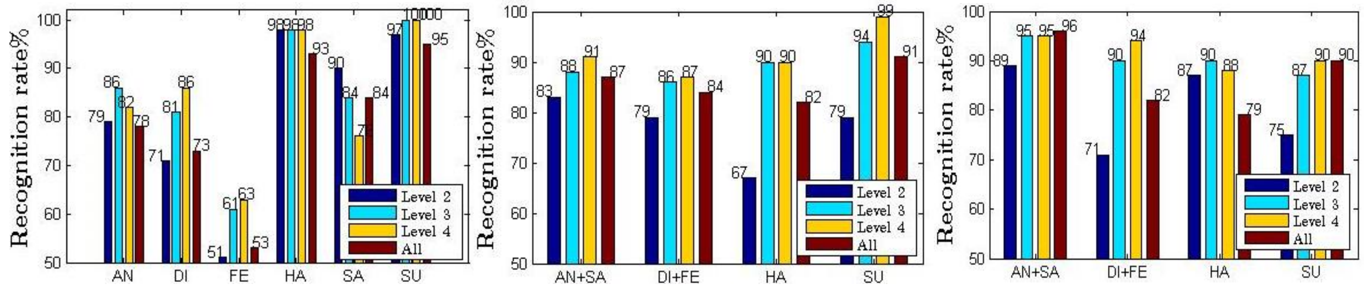
The first experiment examines the influence of the forests' design parameters on the performance of the classifiers. The effect of tree depth was investigated by varying maximum tree depth: 4, 6, 8, 10, 12, 14 and 16, in the training process, with fixed number of trees T = 1000 in the forest. 3D data is used in this experiment. As the forest size is sufficiently large, the results shown in Figure 2 (left) are only from a single trial for each tree depth, as the repeated experiments produce very similar results. It is observed that for supervised learning, smaller trees may not be able to separate the data well. Although the results remain about the same when applying deeper trees - since the random forests are able to handle over-fitting well – large computational resources are required. In the case of unsupervised learning, the recognition accuracy does not strongly depend on depth of trees used in the forests.

The effect of different number of trees in a forest was also tested. The experiments were repeated 10 times for different number of trees (T = 10, 50, 100, 300, 500, 1000) with fixed maximum tree depth of 10 for supervised and 8 for unsupervised approach. The results shown in Figure 2 (right) indicates that having more trees in the forests seen to be beneficial as increasing number of trees helps to get smoother posterior for both methods. This is though at the increased processing time. To achieve desired trade-off between accuracy and computational cost, in the following experiments depths 12 and 8 for supervised and unsupervised method are set, respectively, and T = 1000 for both.

4.2 Supervised vs Unsupervised

In this set of experiments, the tests start on the 3D data in order to compare the performance of the proposed methods employed in supervised and unsupervised learning. The experiments were performed using Matlab on a workstation with an Intel I7-3770S CPU 3.1GHz processor and 8Gb RAM. The average processing times for each face are 0.065s and 0.107s, respectively.

The confusion matrices for supervised and unsupervised



(a) (b) (c)

Figure 3. Recognition rate (%) for (a) the supervised method with 6 classes (b) supervised method with 4 classes, and (c) unsupervised method with 4 classes as a function of different expressions and different expressions' intensities.

		AN	DI	FE	HA	SA	SU	Overall
3D		78.74	75.27	56.90	92.87	81.92	95.21	**80.16**
2D	**Frontal view**	75.59	76.87	52.99	92.82	72.69	95.18	**77.69**
	3 yaw, 3 pitch angles	77.25	78.66	53.88	93.60	70.93	95.81	**78.35**
	5 yaw angles	79.93	78.15	47.48	96.07	69.33	96.59	**77.93**
	5 pitch angles	78.59	77.19	46.22	94.74	67.78	97.19	**76.95**

		AN+SA	DI+FE	HA	SU	Overall
3D		80.78	81.21	80.33	94.45	**84.09**
2D	**Frontal view**	84.66	70.19	88.62	87.98	**81.08**
	3 yaw, 3 pitch angles	84.14	65.57	88.17	87.81	**79.34**
	5 yaw angles	86.30	67.78	89.33	87.63	**80.85**
	5 pitch angles	85.22	68.48	88.74	87.04	**80.53**

Table 3 Comparison of 2D and 3D data using supervised (Top) and unsupervised learning (Bottom).

learning are reported in Table 2. The average recognition rates are 79.31% and 85.93%, respectively. All expression intensities are used in the supervised method, whereas the expressions with the lowest intensity (Level 1) are not used for unsupervised learning, since they are very close to the neutral expression. Additionally, due to a very similar facial deformation, anger and sadness as well as disgust and fear are grouped into the same clusters as they are likely to be confused at a lower expression levels [16] (especially when the class labels are not given in the learning process). Figure 3 summarises the recognition rates for different expressions and expression intensities. On average, higher intensities achieve better performance. For comparison this figure also shows the results obtained for the supervised method with the same data grouping as used for the unsupervised method. In that case the supervised method performs slightly better with the 87.53% average recognition rate.

The recognition performance could be affected by the dimensionality of the reduced space *d*. The next test examined the relation between manifold dimensionality and the recognition rate. The average recognition rate and the standard deviation are tested based on 10 random trials with various dimensionalities (*d* = 2, 3, 4, 5, 7, 10, 20) of the reduced space. As observed in Figure 4, it seems that the average recognition rate and stability of the results are better when the embedded dimensionality is relatively low (blue line). To further investigate the causes of this worsening performance with increased dimensionality of the reduced space, the true class information was used for initialisations of the GMM clustering. In this case the average recognition rate (red line in Figure 4) has slightly decreasing when increase the dimensionality, as the data distribution may not be Gaussian in relatively higher dimensions. It indicates that the "correct" convergence of GMM clustering depends strongly on initialisation in the higher dimensional spaces.

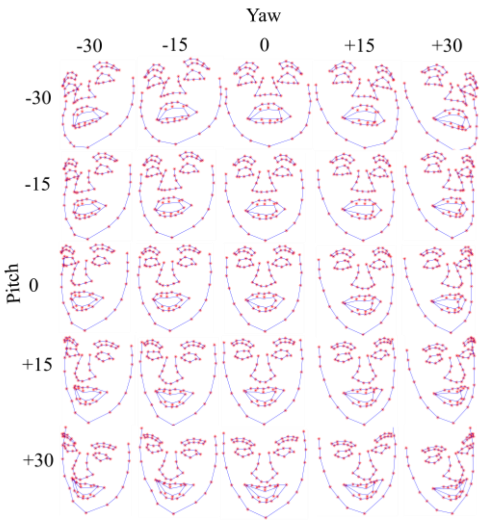


Figure 5. 2D faces projected from 3D data with different yaw and pitch angles.

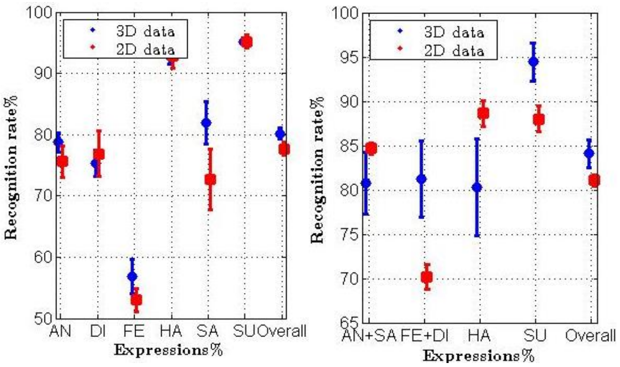


Figure 6. Multiple rounds cross validation results in supervised (left) and unsupervised (right) methods based on 3D and 2D frontal view data.

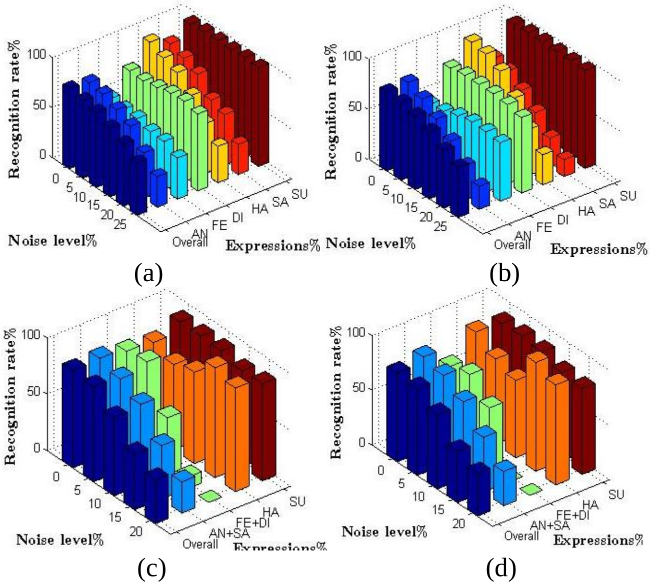


Figure 7. Recognition accuracy as function of the measurement noise level. (a) 3D data using supervised learning. (b) 2D data using supervised learning. (c) 3D data using unsupervised learning. (d) 2D data using unsupervised learning.

4.3 3D data vs 2D data

It is commonly assumed that the use of 3D data can considerably improve the facial expression recognition, since depth information may help to achieve higher sensitivity and specificity when compared to using 2D data only. This assumption is though rarely tested quantitatively. The purpose of this experiment is to compare the performance of the proposed methods when used with 2D and 3D data.

First of all, it should be pointed out that the outcome of the experiments could be significantly influenced by selection of datasets for 2D and 3D analysis. In principle these datasets should be of similar quality, and preferably acquired at the same time. To provide fair comparisons, 2D data is directly projected from the 3D data with various rotation angles. The 2D features are generated from the BU-3DFE database by projecting the 3D landmark feature points with 5 yaw rotation angles (0, ± 15, ± 30) and 5 pitch rotation angles (0, ± 15, ± 30). Figure 5 shows an example of 2D faces of a subject with happiness expression projected from 3D data in various yaw and pitch angles. Numbers of different representative data configurations are used: 2D frontal view faces only; 5 pitch rotation angles without rotation on horizontal direction; 5 yaw angles without rotation on vertical direction; the combination of 3 yaw and 3 pitch rotations (0, ± 15).

To test stability of the methods, multiple rounds of cross-validation using different subset of data are performed. The recognition rate for different facial expressions based on 3D data and 2D frontal view data using both supervised and unsupervised learning are shown in Figure 6. The reported 2D data results are obtained from the projections by the combination of 3 yaw and pitch rotations, as the data projected from other rotation angles achieve very similar results. The averaged results over all rounds are summarised in Table 3. It can be observed from these tables that the use of 3D data always produces slightly better overall results than 2D data. In the supervised method, apart from AN, FE and SA, the improvements achieved for other facial expressions are not significant. Unexpectedly, for the unsupervised method and the AN+SA and HA expressions the recognition on 2D data outperforms the recognition based on 3D data.

4.4 Sensitivity to noise and missing data

Issues like pose, shadows, illumination, etc. could strongly affect the classification performance and therefore the results would be heavily depended on the database used. The use of the simple landmarks make it possible to replace these difficult to control "environmental" influences with the effects these "environmental" aspects have on the detected landmarks which are easier to control and model as these can be robustly and systematically simulated. To facilitate this, along the Gaussian noise, the missing data is also introduced to analyse effects of self-occlusion as well as shadows and illumination changes - as in the context of facial expression recognition with simple landmarks, the outlier problem could be often replaced by the missing data problem.

The first set of experiments is designed to test the impact of noise present in the 3D and 2D data on the performance of the supervised and unsupervised classification. In these experiments, each face landmark position is perturbed by the additive Gaussian noise. The tests are conducted with 5 different levels of noise, which are set to 5%, 10%, 15%, 20% and 25%. Figure 7 (a-b) illustrates the results for the supervised classification using 3D and 2D data, respectively, where 2D data is obtained by projecting using the combination of 3 yaw and pitch rotations. Following the same experimental setup figure 7 (c-d) shows the results obtained for the unsupervised learning. As observed in the figure, the noise does not significantly affect results for some of the expressions, such as SU, DI and FE in the supervised learning. Similarly the results for SU and HA expressions are not significantly affected for the unsupervised classification. Overall, as demonstrated, both the 3D and 2D based recognition are affected by the noise in a rather similar manner.

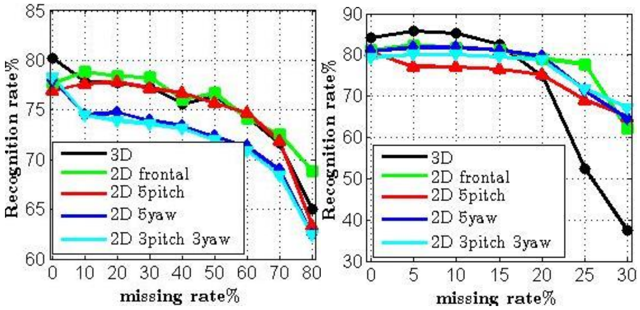


Figure 8. Recognition accuracy on 3D and 2D (frontal view) data in supervised (left) and unsupervised (right) learning of missing data.

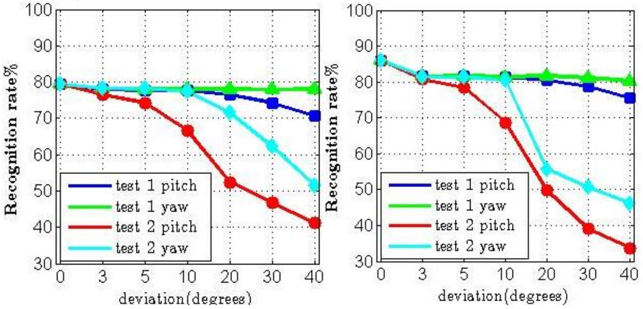


Figure 9. Effect on deviated viewing angles in supervised (left) and unsupervised (right) learning.

In the second set of experiments the effects of the missing data are investigated. To simulate the missing data, up to 80% of the landmarks were randomly discarded in supervised learning, and up to 30% were dropped in unsupervised learning. The recognition rates are tested based on 10 random trials with various rates of missing landmarks. Figure 8 illustrates the average recognition rates as a function of the missing data rate. Since a relatively large number of trees were applied in the experiments ($T = 1000$), the standard deviations are very small, thus they are not shown in the figure. As it can be seen from that figure, the performance remains acceptable with 20% of missing data (or even higher for the supervised learning) for the 3D and 2D data irrespectively of the small head pose changes.

4.5 Varied head poses

The often reported reluctance to use 2D data is based on a belief that the inaccurately estimated head pose may very strongly affect recognition results. As it is often pointed out in literature, small changes in the facial pose can significantly reduce the 2D based recognition accuracy [2]. However, this assertion has rarely been quantitatively tested. For all the experiments, so far described in this paper, is has been assumed that the training set is representative of the possible different head poses in the test dataset. However, collecting the training data for all possible different head poses is not feasible in practice, also robustly estimating the head pose orientation from 2D data is still a challenge [3].

The set of experiments described in this section, is designed to investigate the effects of varied viewing angles on the recognition results when these varied head poses are not represented in the training set. The tests consider two scenarios, for which all the training data are generated by the frontal view projection only. In the first test (test 1), it is assumed that all the test expression faces have a varied head pose but that pose is the same as the pose of their corresponding neutral face. This effectively assumes that although the head pose is unknown it does not change between different expressions. Such assumption may not be realistic for all possible applications. Therefore the second test (test 2) where the head is freely rotated is also conducted. In that experiment the neutral faces are available only with frontal view projection and the corresponding expression faces are acquired with varied head poses. That is, the features for a subject are the distances between all the landmarks of testing faces (possibly non-frontal view) and the corresponding neutral faces (frontal view only). This test is to simulate the case when the pose of head changes between different expressions. In practice, it is unavoidable to have small errors of head pose estimation. This test is to validate whether the method using 2D data is able to cope with these errors.

Figure 9 shows the effect of varied viewing angles on the supervised and unsupervised classification results. The yaw and pitch viewing angles are being changed independently to enable a direct comparison with the results reported in [11]. The results show that in test 1, the recognition rate does not strongly depend on the changing viewing angles. This indicates that when subject does not change the head pose during the face articulation, the results of facial expression recognition are not strongly affected by the unknown head pose. In test 2, although the accuracy falls when the head pose variation exceeds 10 degrees, the results are acceptable for the variations of up to 5 degrees. For the supervised learning the results reported here compare well with results obtained for some complex features. They are very similar to the results of the Topographic Context method proposed in [11] and significantly outperform the Gabor wavelet approach for which test results are also reported in [11]. The analysis of the unsupervised learning was not included in [11]. Overall, the results illustrate that even for the uncontrolled head pose it is still possible to correctly recognize expressions from the 2D data.

5. Discussion

The proposed methods have been quantitatively evaluated using the BU-3DFE database in various situations in order to answer the questions described in Section 2. Through an extensive evaluation it can be concluded that recognition system using only simple landmark features, is able to achieve acceptable recognition accuracy. Although the results produced by applying dense and more sophisticated features (or selecting more discriminative feature points) could be superior, the use of simple features may be important for real-time applications run on low-cost portable devices, as calculation of more complex features may require significantly more computational resources.

In general, the use of 3D information for facial expression improves performance when compared to using 2D information only. This is as expected, since depth information is included in 3D data. The improved recognition rate was observed for some expressions, such as fear and sadness which reflect negative emotion, but not significant improvement was observed for other expressions. It is worth noticing that due to complexity of data collection, 3D data may not be always available. In such cases, using 2D data can still provide acceptable results.

The collection of the labelled training data is a time consuming and expensive task, prone to mistakes possibly leading to unreliable labels. It is therefore useful to consider approaches which do not require such data. In the paper it has been shown that by simplifying the problem, by grouping some of the expression together, it is possible for an unsupervised system to obtain similar recognition performance to a supervised facial expression recognition system.

Based on the reported results obtained for varied head poses, if the head pose does not change during face articulation the result is not dependent on the unknown head pose and therefore the recognition rate is not affected even the pose has not been seen in training set. In the case of the freely moving head the system can still handle small pose variations.

6. Conclusions and Perspectives

In this paper random forest based approaches are presented that recognise the prototypical expressions only using very simple landmark features. The paper shows the possibility of using unlabelled training data for facial expression recognition, and quantitatively investigates the effect of analysing the facial events from 3D and 2D information. It is not claim that the random forest classifier with simple features is better than current state-of-the-art methods which mainly focus on extracting complex features. The important aspect of the paper is to show how well facial expressions with simple features can operate under different conditions, including using 2D data with unknown head pose and unlabelled training data.

The paper has discussed the use of decision forests in both supervised and unsupervised scenarios. But it is very likely in many real scenarios that only a small set of data are labelled with a large set of unlabelled data. Hence a semi-supervised classification would be considered in future research, including dynamic data sets.

Acknowledgement

This work has been partially supported by the SEMEiotic Oriented Technology for Individuals CardiOmetabolic risk self-assessmeNt and Selfmonitoring (SEMEOTICONS) project funded by the FP7 programme (contract number 611516). Authors also would like to thank Dr. Lijun Yin of Binghamton University who has provided the 3D Facial Expression Database (BU-3DFE).

REFERENCES

[1] F. De la Torre and J. Cohn, “Facial expression analysis,” in *Visual Analysis of Humans*. Springer, 2011, pp. 377–409.

[2] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, “Static and dynamic 3d facial expression recognition: A comprehensive survey,” *Image and Vision Computing*, pp. 683–697, 2012.

[3] K. Yurtkan and H. Demirel, “Feature selection for improved 3d facial expression recognition,” *Pattern Recognition Letters*, vol. 38, pp. 26–33, 2014.

[4] P. Liu, J. Zhou, I. Tsang, Z. Meng, S. Han, and Y. Tong, “Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis,” in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 151–166.

[5] F. Zhou, F. De la Torre, and J. Cohn, “Unsupervised discovery of facial events,” in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2574–2581.

[6] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, “Towards practical smile detection,” in *Transactions on Pattern Analysis and machine Intelligence.* IEEE Vol.31(11), 2009, pp. 2106-2111.

[7] A. Criminisi and J. Shotton, *Decision forests for computer vision and medical image analysis*. Springer, 2013.

[8] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic, “Local evidence aggregation for regression-based facial point detection,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 5, pp. 1149–1163, 2013.

[9] http://www.visagetechnologies.com/

[10] X.Zhang, et al. "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database." Image and Vision Computing 32.10, pp.692-706, 2014

[11] J. Wang, L. Yin, X. Wei, and Y. Sun, “3d facial expression recognition based on primitive surface feature distribution,” in Computer Vision and Pattern Recognition, vol. 2. IEEE, 2006, pp. 1399–1406

[12] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, “A 3d facial expression database for facial behavior research,” in *Automatic face and gesture recognition*. IEEE, 2006, pp. 211–216.

[13] R. R. Coifman and S. Lafon, “Diffusion maps,” *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[14] P. Arias, G. Randall, and G. Sapiro, “Connecting the out-of-sample and pre-image problems in kernel methods,” in *Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[15] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[16] P. Ekman and W. Friesen, “Constants across cultures in the face and emotion,” *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

AUTHOR BIOS

Lili Tao is working as a post-doctoral researcher for the SPHERE IRC at the University of Bristol since 2013. Prior to this, she received a B.Eng. in Digital Signal and Image Processing in 2010 with a first class honours degree and a Ph.D in Computer Vision in 2014, both from the University of Central Lancashire, UK. Her research area is computer vision and her research interests include human motion analysis, 3D deformable objects reconstruction, and facial expression analysis. Contact her at lili.tao@bristol.ac.uk

Bogdan Matuszewski is a professor of computer vision in the College of Science and Technology at the University of Central Lancashire, and the head of robotics and computer vision research laboratory at the School of Engineering. His research interests include: use of Bayesian methodology for modelling, tracking and recognition; deformable models, variational and PDE based methods for image analysis applied to data registration, segmentation and interpretation; biomedical and industrial applications of computer vision and machine learning. Contact him at bmatuszewski1@uclan.ac.uk