Deep Features-based Expression-Invariant Tied Factor Analysis for Emotion Recognition

Sarasi Munasinghe, Clinton Fookes, Sridha Sridharan
Image and Video Research Lab, Queensland University of Technology
2 George Street, GPO Box 2434, Brisbane, QLD 4001, Australia
s.munasinghekankanamge@hdr.qut.edu.au, c.fookes@qut.edu.au, s.sridharan@qut.edu.au

Abstract

Video-based facial expression recognition is an open research challenge not solved by the current state-of-the-art. On the other hand, static image based emotion recognition is highly important when videos are not available and human emotions need to be determined from a single shot only. This paper proposes sequential-based and image-based tied factor analysis frameworks with a deep network that simultaneously addresses these two problems. For video-based data, we first extract deep convolutional temporal appearance features from image sequences and then these features are fed into a generative model that constructs a low-dimensional observed space for all individuals, depending on the facial expression sequences. After learning the sequential expression components of the transition matrices among the expression manifolds, we use a Gaussian probabilistic approach to design an efficient classifier for temporal facial expression recognition. Furthermore, we analyse the utility of proposed video-based methods for image-based emotion recognition learning static tied factor analysis parameters. Meanwhile, this model can be used to predict the expressive face image sequences from given neutral faces. Recognition results achieved on three public benchmark databases: CK+, JAFFE, and FER2013, clearly indicate our approach achieves effective performance over the current techniques of handling sequential and static facial expression variations.

1. Introduction

The last two decades have seen an escalating interest in methods for automating the coding of facial expression. Such systems will have numerous applications in a wide range of fields including business; security; consumer applications; education; mental and physical health; automotive and robotics applications. Changes in facial expression are also of significant interest to the biometric community as face based human identification should be robust to expression variations of the subject. Yet, despite this keen interest, the reality is that the promise of computer vision systems to efficiently and accurately recognise facial expressions in natural and open settings has not been achieved. The existing facial expression recognition algorithms can be separated into two categories: image-based and video-based methods. Temporal information of facial expressions is highly important since expressions evolve as a space-time visual phenomena and the temporal domain conveys more information than the static domain. This additional information should be exploited in video based expression recognition methods. Image-based methods assume importance when the the emotion needs to be accurately estimated from a single shot or a few static frames in sequence and this may at times be a harder problem as the significant temporal information is not available.

To combat the challenges of video-based facial expression recognition we propose in this paper novel techniques to express the sequential nature of facial variations in a productive way. Recent studies [6, 9, 28] have shown that the deep neural network (DNN) and the deep convolutional neural network (DCNN) can be used to obtain a more effective feature representation from raw video data [28]. As benchmark video-based facial expression databases such as the Cohn-Kanade Plus (CK+) have a small amount of data, a standalone deep network is not enough for achieving high accuracy due to the overfitting problem. In order to overcome this problem and boost the performance of the video-based facial expression recognition, first we extract the sequential appearance features from image sequences and then these features are fed into the new interpretation of the seminal factor analysis concept across different facial expression sequences. In our work, we are interested in sequential factor analysis since the low dimensional subspace and the manifold learning concepts can cater for significant facial variations [1, 4]; but these have not yet addressed video-based (sequential) facial expression recognition. In this paper, we propose a sequential factor analysis approach to recognize facial expression sequences in video data with extracted sequential appearance features using a deep network.

The sequential expression-invariant Tied Factor Analysis (sequential expression-invariant TFA) generative approach we propose here models the human face as a combination of two factors: an identity factor and a sequential expression-dependent factor. We define a linear model equation to compute and learn all components of our model. First, we implement our generative model based on image-data, and then we extend that static model into the temporal domain, adding sequential components as our goal is to develop a video-based framework. According to our linear process, all components go through two different subspaces: an identity space and an observed space. The identity space is associated with identity variables and the observed space is associated with all sequential expression-dependent parameters based on the video aspect of facial expressions.

As an initial stage, our DCNN is trained using image sequences and the feature vectors of the last convolution layer are taken as a feature representation which is then fed as feature vectors into the sequential expression-invariant TFA model for classification purposes. Then we implement an algorithm to derive all sequential expression-dependent parameters during the training stage. Our algorithm is based on the Expectation-Maximization (EM) technique [15]. In this algorithm, the identity values are estimated to optimize the values of the sequential expression-dependent parameters iteratively. After learning the sequential expression components of an image sequence during the training stage, video-based expression recognition is obtained. In the testing stage, we use a probabilistic approach to obtain the recognition decision of unknown expressive image sequences. Furthermore, this model can be used to predict the expressive face image sequences from given neutral faces. Meanwhile, our method can also be applied for successfully handling image-based facial variations.

In summary, we make the following significant contributions in this paper: (1) We propose a novel sequential expression-invariant TFA generative model using extracted sequential appearance features with a DCNN, in order to recognize video-based facial expressions; (2) We show that our technique can be applied to estimate facial expressions of still images and demonstrate the better performance on spontaneous expressions for the FER2013 wild data set.

2. Related Works

2.1. Facial expression recognition

One of the main challenges of facial expression recognition algorithms is to determine the parameters which are independent from the identity of human faces. In reality, facial expressions can present a great degree of inconsistency in shape and texture. Various studies are proposed to automatically recognize facial expressions.

Patil *et al.* [13] proposed a face expression recognition algorithm for image sequences using an Active Shape Model (ASM) and the Support Vector Machine (SVM). The major drawback of this model is the construction of a strong and complex training process to obtain the trained matrix, which is based on the expressions and recognition. Furthermore, the ASM only locates the shape of the modeled objects, and disregards the texture, so in practice this approach does not take full advantage of the information available. Chumkamon and Hayashi [3] introduced a facial expression recognition framework that uses Constrained Local Models (CLMs) to extract facial features and follows Hidden Markov Models (HMMs) to classify facial expressions. This is the first study that applies the combination of the CLM and HMM concepts to recognize human emotions. They achieved good recognition performance when they executed their framework for a long time with more states, since there are many processes to compute of the facial expression recognition transition. However, they have to expand the current framework in order to obtain simultaneous identity and expression recognition.

Shan *et al.* [17] introduced a model based on the LBP features of face images. They used several machine learning methods to recognize facial expressions. Their experiments showed that the LBP extraction method was more robust than the Gabor wavelet method. They further expanded their model with Boosted-LBP to extract the most important facial features. Finally, they concluded that the best recognition rate is achieved by using a SVM classifier with Boosted-LBP facial features. Moreover, this model works well for low resolution images.

2.2. Low dimensional subspace and factor analysis approaches

In this section, we discuss the most recent work on the factor analysis concept which is based on a low-dimensional subspace. Factor analysis constructs a low dimensional subspace from a multi-scale and high-dimensional identity subspace. The observed subspace represents the facial variations of a person. The identity subspace is stable over all facial variations and is dependent on the human face. In [14], a pose-invariant TFA probabilistic framework is proposed for viewpoint changes. According to their work, the factors depend on the pose variation, but the factor loadings (tied) are defined for each individual separately. Comparable works are presented in [19] and [20]. Both studies use the manifold learning technique to define a mapping between data spaces. These probabilistic frameworks proposed a class dependent factor analysis method to handle facial variations. In [19], a class dependent model is introduced for illumination changes. A probabilistic framework is proposed for both face identification and verification in [20]. Gong *et al.* [5] proposed a new approach for age-invariant face recognition based on the hidden factor analysis concept. Their model is based on two factors: an identity factor and age factor. The identity component is dependent on the face image of the person and the age factor is based on the aging process. Extensive experiments confirmed that their model is superior to the current state-of-the-art age-invariant face recognition algorithms.

In our work, we propose a sequential expression-dependent TFA to explicitly map facial expression sequences and show that the low dimensional subspace and the manifold learning concept can provide remarkable performance over the state-of-the-art.

2.3. Deep neural network models for facial expression recognition

Liu *et al.* [10] proposed a novel Boosted Deep Belief Network (BDBN) for facial expression recognition using three training stages iteratively in a unified loopy framework. In their experiments, they selected the first frame (with neutral expression), and the last three frames from each image sequence, in order to obtain more samples from the CK+ database. Extensive experiments with CK+ and the Japanese Female Facial Expression (JAFFE) databases proved that their framework achieved dramatic improvements over current state-of-the-art algorithms which have been benchmarked on these two databases.

In recent studies, Jung *et al.* [7] and Khorrami *et al.* [8] applied convolutional neural networks (CNNs) to facial expression recognition. In [7], the authors trained a network for facial expression recognition, extracting and combining both appearance and geometric features. Further, they trained their system for video data and automatic facial action detection is another main contribution of their work. In [8], they followed expression recognition for a single image introducing a new approach to decipher which parts of the face affect the classification task.

In contrast, our work is based on facial expression recognition of image sequences that incorporate deep learning-based temporal appearance features.

Recent facial expression models [11, 12, 21, 26] can be considered as a black box, coupled with deep architecture with different configurations. In [25], authors deal with static expression recognition for the Emotion Recognition in the wild challenge using multiple CNN models while in [18], they boosted the wild facial expression recognition performance by replacing softmax classification function with SVM. Both spontaneous facial expression recognition methods generated state-of-the-art performance on the FER2013 database.

3. Expression-Invariant TFA Model

3.1. Image-based Expression-Invariant TFA Model

Initially, we define a TFA generative model to handle image-based facial expression variations. The model leverages the strengths of the standard factor analysis approach including an identity-dependent component but unlocks significant new capability with an expression-invariant representation. Since, observed expressive images are generated by the low dimensional factors in the observed space and image identities are computed by the noisy high dimensional pixels in the identity space, these spaces are referred as a low dimensional observed space and a high-dimensional identity space respectively. A Bayesian generative model and posterior probability distribution are the key concepts behind this method. The generative model can be organized into two key stages: (i) represent *a low-dimensional observed space*, (ii) map a high-dimensional identity data space to a low-dimensional observed space by using *an expression-dependent transformation function*. The model begins with two data spaces: a high-dimensional data space (identity space) and a low-dimensional observed space. In order to achieve a low-dimensional observed space from the high-dimensional identity data space, we employ a Bayesian generative technique.

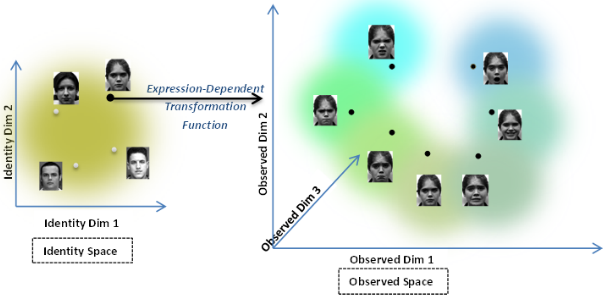

Figure 1. Mapping between identity space and observed emotional space.

Each position in the identity data space indicates a different individual. The identity data space consists of variables which are mapped with each individual. These variables are known as **identity variables**. Each identity variable in the identity data space defines a particular person as shown in Figure 1. This paper further addresses the identity variable of the individual in a multi-dimensional space. Hence we can term it as a multi-dimensional identity variable. Moreover, each position in the low-dimensional observed space specifies a different image variable with facial expressions. The same person will be represented at different places in the observed space, depending on the facial expression of the person as shown in Figure 1. In this TFA generative model, information in the observed space is defined by an **expression-dependent transformation function** as,

$$x_{ije} = m_e + F_e h_i + \varepsilon_{ije}. \tag{1}$$

where x_{ije} represents the *j* image of individual *i* in the e^{th} expression, F_e is a linear function specialized for each expression *e*, h_i is the identity variable of individual *i*, m_e represents the mean of expression *e*, and ε_{ije} is the observed noise. The factors (F_e and m_e) depend on the expression and the factor loadings h_i are the same at each expression. From this expression-invariant TFA model, our goal is to determine all parameters and identify the e^{th} value that depends on expression. Next we extend the above static model into a temporal model as explained in the following section.

3.2. Sequential Expression-Invariant TFA Model for Video-based data

We propose a new sequential model to handle sequential information of facial expressions, learning sequential TFA parameters. The sequential principle is highly important since facial expressions are space-time visual phenomena. Moreover, video sequences capture a wider range of facial expressions as they consist of a large number of sequential images and the temporal domain conveys more information regarding an expression than what is contained in a static image frame.

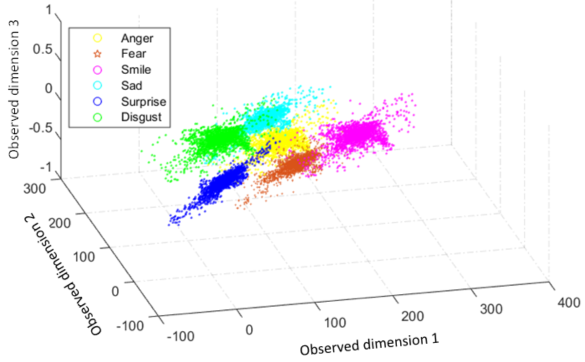

Figure 2. Manifold representation of temporal domain facial expression variation.

Our sequential domain framework consists of a collection of expression manifolds as shown in Figure 2. The transition matrices among the expression manifolds are learned from training videos. For that, we implement TFA models in order to represent the transition matrices among the expression manifolds. Each TFA model (transition matrix) explains the difference between any two consecutive video frames (moving from one expression to another expression). So, we use the sequence of TFA models (sequence of transition matrices) to identify the expression for each input video. Each identity variable in the identity data space defines a particular person as explained in the image-based model. The same person will be represented at different places in the observed space, depending on the video frame of the particular facial expression sequence.

3.3. Learning Sequential Parameters

Model parameters for a facial expression sequence can be formulated as follows,

$$F_e = ({F_e}^1, {F_e}^2, ..., {F_e}^{N-1}), \tag{2}$$

$$m_e = ({m_e}^1, {m_e}^2, ..., {m_e}^{N-1}), \tag{3}$$

$$\Sigma_e = ({\Sigma_e}^1, {\Sigma_e}^2, ..., {\Sigma_e}^{N-1}), \tag{4}$$

where N is the number of the frames. Hence we can reinterpret the earlier expression-dependent transformation function (Equation 1) as follows for video-based data,

$${x_{ije}}^n = {m_e}^n + {F_e}^n h_i + {\varepsilon_{ije}}^n, n = 1, 2, ..., N-1. \tag{5}$$

Here, ${F_e}^n$, ${m_e}^n$, and ${\Sigma_e}^n$ are expression-dependent parameter values between any two consecutive video frames and F_e, m_e, and Σ_e are all expression-dependent parameter values for a particular facial expression sequence.

The relationship between the identity variable of the identity space and the expression-dependent parameters of the observed space are clearly stated in Equations 1 and 5. Since we employ a probabilistic framework, we cannot define an exact value for the identity variable. Thus, we consider the identity variable has a prior distribution. For any two consecutive video frames, we define the probability distribution of the identity variable based on the first frame, h_i ($p(h_i)$), so we can generate the observed variable of the second frame, ${x_{ije}}^n$ as the conditional probability distribution ($p({x_{ije}}^n|h_i)$). These two definitions are given by (6) and (7),

$$p({x_{ije}}^n|h_i) = \mathcal{G}_x[{F_e}^n h_i + {m_e}^n, {\Sigma_e}^n], \tag{6}$$

$$p(h_i) = \mathcal{G}_h[0, \mathbf{I}], \tag{7}$$

where $\mathcal{G}_a[b, C]$ explains a Gaussian distribution in a with mean b and covariance value C. Equation 7 describes a prior distribution of the identity factor and Equation 6 describes the conditional probability distribution of the observed face image.

In order to estimate the model parameters using the EM algorithm, we maximize the following objective function,

$$p(h_i|{x_{i**}}^n) = \prod_{j=1}^{J} \prod_{e=1}^{E} p({x_{ije}}^n|h_i)p(h_i). \tag{8}$$

The above objective function can be explained as a Gaussian distribution with mean EV_m and covariance EV_c as shown in Equations 9 and 10,

$$EV_m[h_i|x_{i**}{}^n] = (\mathbf{I} + \sum_{j=1}^{J}\sum_{e=1}^{E} F_e{}^{nT}\Sigma_e{}^{n-1}F_e{}^n)^{-1}\cdot$$

$$\sum_{j=1}^{J}\sum_{e=1}^{E} F_e{}^{nT}\Sigma_e{}^{n-1}(x_{ije}{}^n - m_e{}^n), \quad (9)$$

$$EV_c[h_i h_i^T|x_{i**}{}^n] = (\mathbf{I} + \sum_{j=1}^{J}\sum_{e=1}^{E} F_e{}^{nT}\Sigma_e{}^{n-1}F_e{}^n)^{-1} +$$

$$EV_m[h_i|x_{i**}{}^n] \cdot EV_m[h_i|x_{i**}{}^n]^T, \quad (10)$$

$$\tilde{F_e}^n = [F_e{}^n \quad m_e{}^n], \quad (11)$$

$$hh_i = [h_i^T \quad 1]^T, \quad (12)$$

$$\tilde{F_e}^n = (\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ije}{}^n EV_m[hh_i|x_{i**}{}^n]^T)\cdot$$

$$(\sum_{i=1}^{I}\sum_{j=1}^{J} EV_c[hh_i hh_i^T|x_{i**}{}^n]), \quad (13)$$

$$\Sigma_e{}^n = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}[x_{ije}{}^n x_{ije}{}^{nT} - F_e{}^n EV_m[hh_i|x_{i**}{}^n]$$

$$x_{ije}{}^{nT}]. \quad (14)$$

The EM algorithm iteratively maximizes the likelihood values of the model parameters until convergence.

4. DCNN Feature Learning

Feature learning and feature extraction stages are very important to accurately represent the facial expressions of human face images. Furthermore, these stages make a significant contribution to reduce the classification error and to improve recognition accuracy. Hence, in our experiments we select an effective DCNN architecture to extract deep feature information from expressive face images since deep features have shown to be the most powerful feature representation methods in many computer vision tasks.

In our framework, we follow a DCNN and extract deep features from image sequences, modeling the video (image sequence) as an ordered sequence of frames. In the DCNN model, the first hidden layer h_1 can be defined as follows using the hyperbolic tangent function, $tanh$,

$$h_1 = tanh(W_1 x + b_1), \quad (15)$$

where h_1 is the output from first hidden layer, W_1 is a weight matrix, b_1 is a bias vector and x is an input pattern.

The output feature vector is then computed as follows;

$$h_5 = tanh(W_5 h_4 + b_5), \quad (16)$$

$$h_6 = vector(h_5), \quad (17)$$

where h_5 is the output features from the fifth hidden layer, W_5 is a weight matrix, b_5 is a bias vector and h_4 is the output from the fourth hidden layer. Finally, a deep feature vector h_6 is computed from h_5 feature maps.

Our DCNN architecture consists of an input layer, three convolution layers, two sub-sampling layers, a fully-connected layer, and the softmax output layer. The DCNN extracts facial features of the input pattern. The convolution kernel size is equal to 5 and outputs maps from the first, second and third convolution layers are equal to 32, 64 and 128 respectively.

5. Probabilistic Approach for Recognition

After the deep feature extraction process, each face image is expressed by a deep feature vector. Then we implement the sequential expression-invariant TFA model based on image sequences of the CK+ training data set. As a final stage we classify the expression sequence of the given image sequence as follows. An overview of the entire framework is illustrated in Figure 3.

The given image sequence, I_{seq} can be formulated as below,

$$I_{seq} = (x_1, x_2, x_3, ..., x_N). \quad (18)$$

The trained expression sequence models can be defined as below,

$$M = (M^1, M^2, M^3, ..., M^J). \quad (19)$$

For J = k,

$$M^k = (M^k{}_{1,2}, M^k{}_{2,3}, ..., M^k{}_{S-1,S}), \quad (20)$$

where N = number of frames in the given image sequence, J = number of trained expression sequence models, and S = number of frames in sequence model, M^k.

Our goal is to determine the posterior probability that the given image sequence matches each trained expression sequence model and finally identify the maximum posterior probability model as below,

$$P_k = \max_{1\leq j\leq k}(\prod_{i=1}^{\frac{N}{2}} p(x_{2i-1}, x_{2i}|M^j{}_{2i-1,2i})). \quad (21)$$

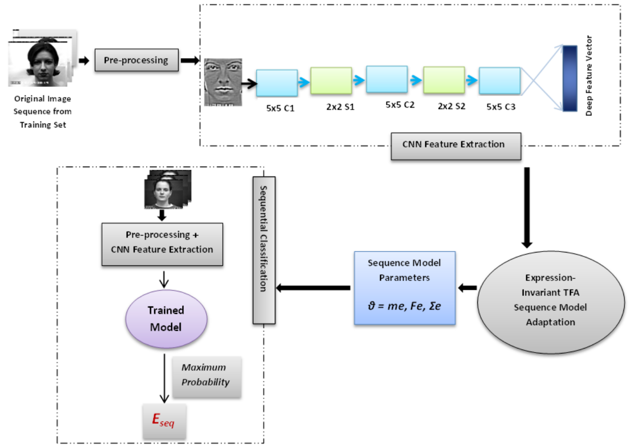
Original Image Sequence from Training Set
Pre-processing
5x5 C1
2x2 S1
5x5 C2
2x2 S2
5x5 C3
Deep Feature Vector
CNN Feature Extraction
Pre-processing + CNN Feature Extraction
Trained Model
Maximum Probability
E_{seq}
Sequential Classification
Sequence Model Parameters
$\vartheta = me, Fe, \Sigma e$
Expression-Invariant TFA Sequence Model Adaptation

Figure 3. Outline of the entire framework. At the training stage, the training images are pre-processed, followed by DCNN feature extraction (Section 4) on each training image. After the DCNN feature extraction process, the sequential expression-invariant model is implemented (Section 3) by the training images (CK+ database). At the testing process, the expression sequence of the given unknown image sequence (Section 5) is obtained by extracting features, and then using the probabilistic approach. Finally, the maximum posterior probability model is computed, in order to achieve the recognition decision.

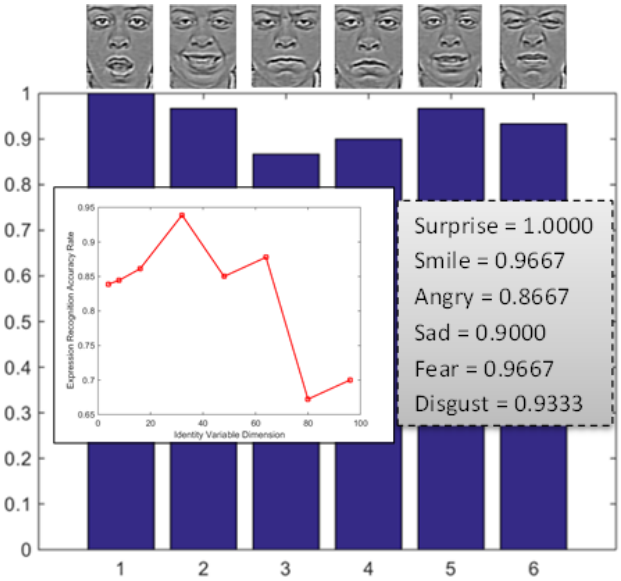

Figure 4. Performance comparison on the CK+ in terms of classification rate for six facial expressions.

6. Experiments

6.1. Image-based static experiments

First, our model is evaluated on CK+ image-based pixel data to determine that the model has identified the relationship between the identity data space and the low dimensional observed space. In image-based experiments, we selected the first frame (neutral expression) and the last frame (peak expression) from each image sequence. During training, we learned the unknown parameter values (expression-dependent linear matrix, mean vector, and covariance vector). Meanwhile we implemented a single expression-invariant TFA model. This expression-invariant TFA model consists of 6 sub models, each including the parameters to explain the difference between one of the 6 expressions (surprise, smile, anger, sadness, fear, and disgust) and the neutral expression. In Figure 4, we present results for six different facial expressions. It is clearly seen that the surprise expression achieves the highest performance compared to other expressions. The overall facial expression recognition performance is 93.9% with 32 identity variable dimensions.

Recognition Method	Accuracy
Ada+SVM(Linear) [17]	0.404
Ada+SVM(Poly) [17]	0.404
Ada+SVM(RBF) [17]	0.413
BDBN [10]	0.680
Proposed Method with image-data	**0.722**

Table 1. Trained on CK+ database and tested on the JAFFE database, in terms of average classification rate.

We also followed the cross-database validation as shown in Table 1: trained on the CK+ database and tested on the JAFFE database. Images of all two databases are processed using illumination normalization process, in order to keep the same lighting condition. Our proposed method achieves 0.722 expression recognition accuracy on the JAFFE database. Cross-database validation results usually shows low performance in the state-of-the-art. However, our method achieves stronger results than the benchmark BDBN approach [10]. Hence, we conclude that our trained model can be easily adjusted to another data set in order to recognize facial expressions.

6.2. Experimental Setup for CK+ video data

For video-based experiments, we used 64×80 resolution images from the CK+ database. The CK+ database has 593 image sequences from 123 subjects. We selected 327 image sequences from 100 subjects that can be labeled as one of six expression sequences: surprise, smile, anger, sadness, fear, and disgust. Our selection criterion was that the video sequence is equal or longer than 7 frames and a sequence can be labelled as one of the six basic expressions (smile, surprise, sad, anger, fear, and disgust). Most previous research has only used the last frame as they contain the peak (apex) of the expression. We follow a segmentation procedure which enables us to overcome the challenging problem of the different lengths of the sequences. In CK+ database, all sequences start from a neutral expression and continue gradually to the apex. We find that we can select 7 significant frames which are able to demonstrate the sequential behaviour of expression sequences with arbitrary length. If the length of the expression sequence is equal to n, then we select the first frame, $n/6$ frame, $n/3$ frame, middle frame, $2n/3$ frame, $5n/6$ frame, and the last frame.

Next we randomly divided all individuals into ten groups and followed a leave-one-group-out cross validation. All images were segmented from the background and processed through an illumination normalization process in order to keep the same lighting condition. During training, we learned the unknown parameter values (sequence of expression-dependent linear matrix, mean vector, and covariance vector). We implemented six sequential expression-invariant TFA models. Each sequential expression-invariant TFA model consists of the parameters to explain the difference between any consecutive frames of the expression sequence. Our sequential expression-invariant TFA model employs the EM algorithm to calculate the unknown sequential expression-dependent parameters.

During the training process, we trained the DCNN on the CK+ training sets, extracting activation neurons from the last hidden layer of the DCNN architecture. We implemented a *32c-2s-64c-2s-128c* DCNN architecture. Then we fed the extracted activation neurons of the last hidden layer of the training data set to the training model and obtained model parameters. We trained our model for *10 iterations* to obtain the optimized values for model parameters. Once the model has been tuned, we used testing data set to measure the recognition performance.

6.3. Experiments with deep learning-based video data

Expression	Accuracy	Expression	Accuracy
Surprise	1.0000	Sadness	0.9500
Smile	0.9834	Fear	1.0000
Anger	0.9334	Disgust	0.9667

Table 2. Performance Comparison on the CK+ in terms of classification rate for six facial expression sequences.

In this section we explain the experiments carried out based on the deep learning-based video data. In Table 2, we present video-based results for six different facial expressions. It is clearly seen that, each expression sequence achieves better performance than static data above. The overall facial expression recognition rate is 97.23% for deep learning-based video data. Table 2 results show that the worst performance is obtained across the anger and sad expression sequences. Furthermore, we compared the performance of our deep learning-based temporal model with recent dynamic facial expression recognition methods for each expression.

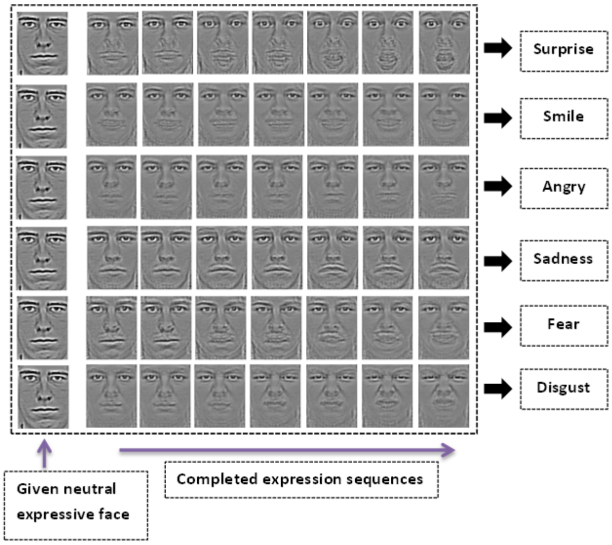

Figure 5. Predicted results from sequential expression-invariant TFA. Column 01: Original neutral image from test set. Column 02 to Column 07: Predicted results for six different expression sequences.

In Figure 5, we visualize the prediction results of six different expressive image sequences from the given neutral face. In this test, we calculated the identity variable value of the given neutral face as a posterior distribution. Then we transformed this probability value of the identity variable to the observation space by using the learned parameters from one of the sequential expression-invariant TFA models (neutral-surprise, neutral-smile, neutral-angry, neutral-sad, neutral-fear, and neutral-disgust). In order to visualize the predicted images, we employ a probability distribution over the predicted images. It is clearly seen that our model can successfully predict the expression sequences, but the identity of the predicted faces contains obvious noise due to the Bayesian nature of our framework. However, the noiseless prediction is not an important factor for recognition decisions, since all the major variations of each expression are captured and stored in the sequential expression-invariant TFA model for each individual.

We also benchmark our performance through a comparison of image-based and video-based facial expression recognition performance of our proposed method with other state-of-the-art expression recognition approaches as shown in Table 3. After analyzing Table 3, we summarize that the average expression recognition results of our method (with deep learning) is the best among temporal expression recognition methods in the state-of-the-art. Furthermore, it is clearly indicated that the combination of DCNN for feature extraction, followed by TFA for classification performs better than either a straight DCNN system with the softmax classifier or the sequential TFA with a pixel representation. According to the individual performance of each component (the individual sequential TFA achieved 94.45% and the individual DCNN achieved 94.83%), we can analyze that each component equally contributes to the final results. However, our method obtains remarkable performance for temporal expression recognition by presenting a single deep network model and a sequential TFA approach with a simple probabilistic classifier. Since the transition matrices among the expression manifolds are effectively highlighted through the sequential TFA expression factors which are based on the deep convolutional features, our proposed model works so well for video-based expression recognition even with a simple classifier.

Hence the proposed framework has a value within the video-based expression recognition community for effectively managing varying expressions through a more optimal combination of a deep hierarchical feature representation managed by a generative classification approach.

Recognition Method	Accuracy (%)
CSPL* [27]	89.9
Combined features + Adaboost* [23]	92.3
AdaGabor* [2]	93.3
Image-based TFA with pixel data*	**93.9**
LBPSVM* [17]	95.1
BDBN* [10]	96.7
Zero-bias CNN+AD* [8]	98.3
One-shot Learning [24]	86.7
Bayesian Temporal Manifold [16]	91.8
LGBP-TOP [22]	92.0
Sequential TFA with pixel data	**94.45**
Straight DCNN with softmax classifier	**94.83**
DTAGN [7]	96.94
Sequential TFA with DCNN features	**97.23**

* indicates the static expression recognition methods

Table 3. Comparison with recent advances in expression recognition on CK+ dataset.

6.4. Experiments with FER2013 database

For experiments in the wild, we used 48×48 resolution images from the FER2013 database. The training set consists of 28,709 images. The public test set consists of 3,589 examples (validation set) and the private test set consists of another 3,589 examples (final test set). During the training process, we trained the unknown wild expression parameter values and implemented seven expression-invariant TFA models (including neutral) based on the wild conditions.

Our system achieved 0.7111 accuracy with the private leaderboard data set and 0.6987 accuracy with the public leaderboard data set which shows comparative performance with against the benchmarked methods. Table 4 represents the performance comparison of our proposed model with multiple network learning [25] and DLSVM [18] approaches. The multiple network learning [25] method obtained the best results on FER2013 wild database and the DLSVM [18] method achieved remarkable performance by replacing the softmax function with the linear SVM function (69.4% for public leaderboard and 71.2% for private leaderboard). Our method shows effective results using a combination of a DCNN (with softmax) and a wild image based TFA model. This is the one of the main contributions of this work.

Dataset	DLSVM [18]	MNL [25]	Ours
Public Validation	0.694	0.7	0.6987
Private final test	0.712	0.72	0.7111

Table 4. Performance comparison with the benchmarking approaches on FER2013.

7. Conclusion

Sequential expression-invariant TFA with a deep network has been shown to handle both video-based expression recognition and prediction tasks through a complete probabilistic framework used for classification, and incorporation of learned sequential model parameters. Our method can also be applied for successfully handling image-based facial variations. Our system performance is very competitive on the benchmarking CK+ database. It achieved 97.23% temporal expression accuracy which is the best performance in the benchmarking exercises on CK+ database. Hence the proposed framework has a value within the video-based expression recognition community for effectively managing varying expressions through a more optimal combination of a deep hierarchical feature representation managed by a generative classification approach. Moreover, our model provides effective and comparable performance for static expression recognition among cross-database evaluation approaches in the state-of-the-art. Furthermore, one of the important contributions of this work is that our model provides remarkable performance with the FER2013 wild static expression database. In future work, we will examine a more complex generative model that is based on the joint factor analysis concept to explore different facial variations. We will also investigate the performance of our approach on more challenging databases: spontaneous facial expression databases such as AFEW and SFEW.

8. Acknowledgment

This research has been supported by an Australian Research Council Discovery Project [DP140100793].

References

[1] M. Bartlett, J. R. Movellan, and T. Sejnowski. Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464, Nov 2002.

[2] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. IEEE, 2005.

[3] S. Chumkamon and E. Hayashi. Facial expression recognition using constrained local models and hidden markov models with consciousness-based architecture. In *System Integration (SII), 2013 IEEE/SICE International Symposium on*, pages 382–387. IEEE, 2013.

[4] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2872–2879, Dec 2013.

[5] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2872–2879. IEEE, 2013.

[6] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.

[7] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim. Deep temporal appearance-geometry network for facial expression recognition. *arXiv preprint arXiv:1503.01532*, 2015.

[8] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[11] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[12] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. Facial expression recognition from world wild web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–65, 2016.

[13] R. A. Patil, V. Sahula, and A. Mandal. Facial expression recognition in image sequences using active shape model and svm. In *Computer Modeling and Simulation (EMS), 2011 Fifth UKSim European Symposium on*, pages 168–173. IEEE, 2011.

[14] S. Prince, J. Warrell, J. Elder, and F. Felisberti. Tied factor analysis for face recognition across large pose differences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):970–984, June 2008.

[15] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

[16] C. Shan, S. Gong, and P. W. McOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *BMVC*, pages 297–306. Citeseer, 2006.

[17] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[18] Y. Tang. Deep learning using support vector machines. *CoRR, abs/1306.0239*, 2, 2013.

[19] B. Tun and M. Gkmen. Manifold learning for face recognition under changing illumination. *Telecommunication Systems*, 47(3-4):185–195, 2011.

[20] B. Tunc, V. Dagli, and M. Gokmen. Robust face recognition with class dependent factor analysis. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6, Oct 2011.

[21] J. Wang and C. Yuan. Facial expression recognition with multi-scale convolution neural network. In *Pacific Rim Conference on Multimedia*, pages 376–385. Springer, 2016.

[22] L. Xie, H. Wei, W. Yang, and K. Zhang. Video-based facial expression recognition using histogram sequence of local gabor binary patterns from three orthogonal planes. In *Control Conference (CCC), 2014 33rd Chinese*, pages 4772–4776. IEEE, 2014.

[23] P. Yang, Q. Liu, and D. N. Metaxas. Exploring facial expressions with compositional features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2638–2644. IEEE, 2010.

[24] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1635–1648, 2013.

[25] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.

[26] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect“in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47, 2016.

[27] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE, 2012.

[28] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.