Computer Vision and Image Understanding 162 (2017) 146–165

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Non-rigid registration based model-free 3D facial expression recognition

Arman Savran[a,1,*], Bülent Sankur[b]

[a] Italian Institute of Technology, Genoa 16163, Italy
[b] Department of Electrical and Electronics Engineering, Bogazici University, Istanbul 34342, Turkey

ARTICLE INFO

Article history:
Received 4 November 2016
Revised 28 April 2017
Accepted 20 July 2017
Available online 22 July 2017

Keywords:
Facial expression recognition
3D face analysis
Model-free
Non-rigid registration
Shift-invariance
Action units

ABSTRACT

We propose a novel feature extraction approach for 3D facial expression recognition by incorporating non-rigid registration in face-model-free analysis, which in turn makes feasible data-driven, i.e., feature-model-free recognition of expressions. The resulting simplicity of feature representation is due to the fact that facial information is adapted to the input faces via shape model-free dense registration, and this provides a dynamic feature extraction mechanism. This approach eliminates the necessity of complex feature representations as required in the case of static feature extraction methods, where the complexity arises from the necessity to model the local context; higher degree of complexity persists in deep feature hierarchies enabled by end-to-end learning on large-scale datasets. Face-model-free recognition implies independence from limitations and biases due to committed face models, bypassing complications of model fitting, and avoiding the burden of manual model construction. We show via information gain maps that non-rigid registration enables extraction of highly informative features, as it provides invariance to local-shifts due to physiognomy (subject invariance) and residual pose misalignments; in addition, it allows estimation of local correspondences of expressions. To maximize the recognition rate, we use the strategy of employing a rich but computationally manageable set of local correspondence structures, and to this effect we propose a framework to optimally select multiple registration references. Our features are re-sampled surface curvature values at individual coordinates which are chosen per expression-class and per reference pair. We show the superior performance of our novel dynamic feature extraction approach on three distinct recognition problems, namely, action unit detection, basic expression recognition, and emotion dimension recognition.

1. Introduction

Human facial expressions convey diverse set of signals which can be associated with mental states such as emotions, with physiological conditions like pain or tiredness, or with various non-verbal social communication messages. There are many potential applications of expression recognition systems. For instance, Cowie et al. (2001) mention about two hundred emotional states and discuss applications in areas from medicine to education and entertainment.

Expression recognition is a challenging problem, not only due to the variety and subtlety of expressions, but also due to the hurdles in the extraction of effective features from facial images. Typically, automatic expression recognition starts with face detection, followed, possibly, by face pose normalization, and then by feature extraction and classification. Expression recognition algorithms must be robust against such confounding factors as variations in illumination, 3D face pose, subject identity, and texture-like facial hair and make-up. Arguably the most crucial component is the feature extraction since features should be resilient, on the one hand, to the effects of these confounding factors; on the other hand, they must capture the critical facial details enabling discrimination of expressions. Moreover, choice of effective features can also simplify the design and training of the classifiers.

3D acquisition of facial expression images can inherently mitigate some of these challenges. Depending on the 3D reconstruction technique, 3D data can be immune to a great range of illumination and texture variations, and it is not as sensitive as 2D light images to yaw and pitch type out-of-plane rotations. Moreover, 2D light images may fail to capture subtle but discriminative changes on the face if they do not cause sufficient luminance

* Corresponding author.
E-mail addresses: arman.savran@boun.edu.tr, arman.savran@iit.it (A. Savran).
[1] This paper is based on the work carried out in Department of Electrical and Electronics Engineering, Bogazici University.

http://dx.doi.org/10.1016/j.cviu.2017.07.005
1077-3142/

changes, such as bulges on the cheeks and protrusion of the lips. In fact, 3D has been shown to achieve better recognition than conventional 2D light cameras for many types, if not all, of facial actions (Savran et al., 2012a). Notice, however, that 3D modality has its own difficulties, and there are niche instances of facial actions where light images perform better than 3D, such as in some action units around the eyes. Moreover, combining 2D texture with 3D modality improves the overall recognition performance (Savran et al., 2012a). However, the aim of this paper is not to investigate the best performing full recognition system (i.e., by combining the best face detection, the best face pose alignment, the best feature extraction, 2D+3D modality fusion, the best classification, etc.). Instead, we conjecture that 3D modality provides a novel more convenient feature extraction approach which is fundamentally different from conventional methods.

Our main conjecture is that 3D *expression* recognition based on estimated deformation field, but without the guidance of any kind of face modeling is possible and it leads to very simple yet effective features. This approach provides both theoretical and practical advantages. Theoretically, one does not need to make any simplifying assumptions, therefore potential biases due to the design of face models are avoided; similarly, one does not need to recur to local feature models (use of local-context), which are typically high dimensional and cumbersome. Current methods in the literature typically perform local patch analysis resulting in multi-dimensional responses at each location, require landmark detectors and more recently big datasets to construct deep feature representations. The practical advantages of the proposed approach compared to face model-driven methods are that it avoids the tedious face model construction stage, which usually requires expertise, and also there is no need for a model fitting stage, with concomitant problems of complexity and sensitivity to fitting errors. Our method has also some practical advantages compared to deep representations in that it does not necessitate collection of huge datasets, data augmentation and pooling for handling local spatial variations, and efforts to design an adequate deep architecture which involves long experimentation durations over big datasets.

To the best of our knowledge, all the prior expression recognition methods based on deformation estimation, whether 3D or conventional, have used models of face shape; for instance, in the case of 3D data, landmarking (Berretti et al., 2010; Fang et al., 2012; Maalej et al., 2011), active shape/appearance models (Sun et al., 2008; Tsalakanidou and Malassiotis, 2010) or morphable models (Mpiperis et al., 2008; Ramanathan et al., 2006); in the case of 2D luminance images, recent landmarking techniques such as constrained local models (Chew et al., 2012; Chu et al., 2017; Eleftheriadis et al., 2015; Zeng et al., 2016; Zhao et al., 2015). Nevertheless, we do not claim definitive performance superiority of our method (face model-free deformation-estimation-based feature extraction) over methods based on face models or deep models (when necessary amount of training data is available). In principle, it may also be possible to achieve high recognition performances with well-engineered facial landmark-guided methods and feature extractors, or with deep models via end-to-end machine learning. However, our work opens up a third path for 3D data, with the merits that it is free from efforts to develop landmark detectors and feature extractors which model the local-context, and it can operate without the need for very large amounts of training samples to construct complex deep models. The simplicity of our feature representation is because facial information is used to adapt to the input faces via shape model-free dense registration providing a dynamic feature extraction mechanism. Consequently this eliminates the necessity of complex feature representations which would be needed in the case of static feature extraction that obtains features always from pre-determined fixed coordinates in a reference coordinate system, i.e., without adapting their locations depending on the actual input face.

Shape model-free image recognition based on estimated deformation field has been applied in the context of handwritten digit/character or medical image recognition tasks (Keysers et al., 2007). In these works involving an image-matching framework, distances for the purpose of nearest-neighbor classification are computed after a mapping has been estimated between the test input and the reference images. In principle, though deformable image-matching can be applied for face recognition purposes, it is not viable for facial expression recognition as this task requires local analysis. With the help of hand-crafted masks of deformation regions and manually chosen reference faces, our previous work has implemented a shape-free deformable method for 3D expression recognition (Savran and Sankur, 2009). However, the assumption that local expression actions have their prototypical region masks limits the set of expressions or isolated facial actions that one can detect. Consequently, this method falls short of meeting the challenge of complex expressions where one encounters high-degree of variations and co-occurrences of local deformations.

We propose a novel framework to realize a face model-free deformation-estimation-based feature extraction for 3D *expression* recognition. Our framework is based on the premise of optimally selected facial coordinates from the domains of also optimally chosen multiple face registration references. In this context optimality is defined as facial coordinates and face references that maximize the recognition rate, and not the registration accuracy, over a training set. Basically, a different set of facial coordinates on selected reference domains are identified in the training stage. Then an input test face is registered to each of the reference faces by deforming the references towards the input, and then the test input is resampled multiple times at the designated coordinates specific to each reference face. We have made the following choices for the implementation of the above framework. We use 2D projections of the 3D facial surfaces for saving in substantial amount of computations (Savran et al., 2012a). We use a simple deformation model for non-rigid registration, as prior work has shown that further complexity does not improve the recognition performance (Keysers et al., 2007) but may make the computations intractable. As for the features, we use mean curvature due to its comparatively superior performance (Savran et al., 2012a; 2012b) as well as due to its compactness (scalar representation of deformation at each point). Since non-rigid registration provides invariance to local transformations, feature models with large local support (like spin images, Gabor, LBP, HoG, SIFT, etc.) become unnecessary. Finally, we utilize boosting as the selection mechanism.

We show quantitatively that the framework consisting of non-rigid registration to multiple-references and optimally selected sampling coordinates on them enables higher amount of information gain in classification of facial actions. Our experiments on various datasets prove that such information gain leads to higher recognition performance. Our approach is generic with respect to *expression*[2] analysis tasks. As a point in case, we demonstrate the performance of our method on the action unit (AU) detection problem, on the basic expression recognition problem, and finally, on the emotion dimension prediction problem. These experiments involve three 3D face databases of different nature and quality.

The rest of the paper is organized as follows. Section 2 gives a brief overview of face model-driven and model-free *expression* recognition as well as shape model-free non-rigid registration literature. The databases used in the experiments are described in

[2] The term *expression* is used here in a broad sense, including facial action units.

Section 3. Section 4 presents the proposed non-rigid registration method. In Section 5, we develop our registration based recognition approach. Section 6 is devoted to experiments and discussions of the results. In Section 7, we further discuss some important consideration regarding to our face model-free framework. Finally, we give the conclusions in Section 8.

2. Prior work

A large variety of features have been used in the literature for facial expression recognition, described in recent surveys such as in Sariyanidi et al. (2015) for conventional light cameras and in Corneanu et al. (2016) and Sandbach et al. (2012b) for 3D cameras. These feature extraction methods, whether applied to still or video images, can be categorized into two fundamental groups, namely, face model-driven and face model-free. While both approaches have their virtues and drawbacks, a comparative discussion is as follows.

2.1. Face model-driven recognition of expressions

Face model-driven implies that a pre-designed model of human faces is fitted to the input facial data before performing any analysis task. The simplest model-based methods focus on detection of a large number of facial landmarks, for example, up to 83 landmarks as in Wang et al. (2006), Soyel and Demirel (2007), Tang and Huang (2008), Berretti et al. (2010) and Maalej et al. (2011). There are also methods that require fewer landmarks, e.g., Fang et al. (2012) utilizes 12. However, although these studies report high recognition performances, they all employ manual landmarks. Precision of automatic landmark detection is not guaranteed to be high enough in practice, and robust landmark detection on 3D faces, especially under expressions, continues to be a challenging task in Creusot et al. (2013).

A popular approach is to restrict the allowable space of the fiducial points to plausible locations and to variations that are learned from real data, like in Active Shape Models, Active Appearance Models, and constrained local models (Chew et al., 2012). Prior works (Sun et al., 2008; Tsalakanidou and Malassiotis, 2010) have directly used such constraint information on luminance data concomitantly captured with 3D depth scans. Joint statistical models of local 3D geometry and texture features have also been constructed for landmarking to recognize *expressions* (Zhao et al., 2013). Notice that, all these 3D works suffer from sensitivity to luminance variations in the model fitting stage. However, landmarking on luminance images is an active research field and new technique like Ren et al. (2014) and Kazemi and Sullivan (2014) may lead to improved expression recognition performances. Some current methods on 2D luminance data which employ landmark-guided feature extraction are Zhao et al. (2015), Eleftheriadis et al. (2015), Zeng et al. (2016) and Chu et al. (2017).

An alternative statistical shape modeling approach is dense modelling of 3D geometry by 3D morphable models (3DMM) (Blanz and Vetter, 1999), which has been applied to 3D expression recognition in Ramanathan et al. (2006), and also with bilinear modeling to account for identity and expression related variations, simultaneously (Mpiperis et al., 2008).

2.2. Face model-free recognition of expressions

Face model-free *expression* analysis does not depend on any prior face shape model. A common approach is to extract a high-dimensional dense feature set from the images, and then apply dimension reduction, such as by feature selection, e.g., via AdaBoost (Gehrig and Ekenel, 2011; Littlewort et al., 2011; Sandbach et al., 2012a; 2012c; Savran et al., 2012a). Current feature techniques are based on either filter-banks or image descriptors. One of the best performing filter-bank methods are Gabor wavelets applied at multiple orientations and scales (Littlewort et al., 2011). Gabor filters have been shown to perform better than independent component analysis and non-negative matrix factorization (Savran et al., 2012a) features. Gehrig and Ekenel (2011) have demonstrated that block-based discrete cosine transform filters work as well.

The descriptor-based techniques typically compute histograms of low-level image features over local patches. The most commonly employed descriptors are Local Binary Patterns (LBPs) (Shan et al., 2009) and its various extensions. Histogram of oriented gradients has also been applied successfully (Dahmane and Meunier, 2011). Recent methods have combined descriptors with filterbank outputs to achieve modest improvements, as in Local Gabor Binary Patterns (Wu et al., 2012) and Local Phase Quantisation (LPQ) (Dhall et al., 2011) where the latter provides some insensitivity to image blur. Alternative face model-free methods use bag-of-words using SIFT descriptors (Sikka et al., 2012) and deep learning architectures (Kahou et al., 2013).

All these methods can be applied to 3D data as well, provided that 3D data is first converted to some geometry map. Examples are Gabor filters on surface curvature map (Savran et al., 2012a), binary pattern analysis on surface normal maps (Sandbach et al., 2012c) and depth maps (Sandbach et al., 2012a), Zernike moments on depth maps (Vretos et al., 2011), and histogram-descriptors on curvatures (Savran et al., 2013).

Recent work on 3D expression recognition has also focused on spatio-temporal feature extraction to be able to benefit from more information available in the temporal context. For instance, Nebula features (Reale et al., 2013) are constructed by combining the histograms which are extracted from fixed blocks over frontal face. The histograms of the shape labels that are obtained by thresholding the principal curvature values are calculated over a spatio-temporal temporal support. Therefore Nebula features are categorized as histogram-descriptors extended to spatio-temporal feature extraction. More methods to exploit the temporal context have been proposed on 2D images compared to 3D expressions (Corneanu et al., 2016; Sariyanidi et al., 2015), like the popular LBP-TOP method (Zhang et al., 2016; Zhao and Pietikainen, 2007), showing that temporal context is helpful in increasing the recognition rate. However, the scope of our work does not involve the temporal feature extraction aspect.

State-of-the-art in facial expression recognition, as in many recognition problems, is the deep learning approach, which performs end-to-end learning starting from feature extraction with convolutional neural network (CNN) layers. By nature these models are face model-free. Effectively exploiting very big scale datasets, deep learning enables complex hierarchical feature representation. Effectiveness of deep learning has been shown on generic object recognition (Krizhevsky et al., 2012; Szegedy et al., 2013), and they can achieve impressive performances on large scale face recognition problem (Parkhi et al., 2015; Sun et al., 2014). Consequently the success of deep learning has shifted the state-of-the-art in facial expression recognition (Kahou et al., 2013; Ding et al., 2016; Jaiswal and Valstar, 2016; Jung et al., 2015; Khorrami et al., 2015; Kim et al., 2015; Levi, 2015; Ng et al., 2015; Zhao et al., 2016) from engineered local feature extractors (Gabor, LBP, HoG, SIFT, etc.) as well. We want to notice that, before deep-learning became the state-of-the-art, the latest studies on 2D facial expression recognition employed facial landmarks detectors, i.e., for the guidance of face shape models for feature extraction, to better adapt the non-rigidity of faces. Landmarks are especially of help for local AUs; for instance, Eleftheriadis et al. (2015) extract LBP features or Zhao et al. (2015) extract SIFT features around 49 fiducial facial points.

Moreover, landmark detection has recently been employed to facilitate the deep learning by processing only the potentially more relevant regions (Jaiswal and Valstar, 2016). However, use of landmarks violates the end-to-end learning as well as face model-free philosophy.

The first difficulty in deep learning is the need for very large-scale labeled training datasets. However, obtaining accurate facial emotion labels is a very time consuming task, which gets even more complicated and time consuming when the task is to label facial action units. Due to the absence of sufficiently large labelled facial expression datasets for effective training of deep models, current deep-learning models apply fine tuning or transfer learning. For instance, Levi (2015) first train on CASIA Webface (Yi et al., 2014) images which involve 500K face images; however, these authors had to apply LBP feature representation to reduce confounding variations to achieve high recognition rates for basic emotion classification task (7 classes) and had to do fine-tuning of the deep model on the expression dataset. To be able to obtain an effective deep model on small datasets without requiring hand-crafted feature extractor in the first layer, Ng et al. (2015) follow deep CNN transfer learning approach by employing pre-trained network, which is itself trained on the generic ImageNet database (Krizhevsky et al., 2012) involving 1.2M images. They propose a supervised fine-tuning approach which involves two stages: the first stage performs fine-tuning on a big facial expression dataset with around 30K labeled faces (FER-2013 database), and the second stage continues fine-tuning on a small dataset with 1K labeled faces (EmotiW dataset). Very recently Ding et al. (2016) proposed an alternative approach to handle small datasets based on regularization of a deep face recognition net for the expression recognition task. Their feature level regularization exploits the rich facial information in the facenet (Parkhi et al., 2015) which is trained on 2.6M face images. In all these above methods, first, face pose alignment is performed by facial landmark detectors, and then data is further augmented by random cropping as well as by vertical mirroring. This procedure, which typically increases the sample-size by more than an order of magnitude, becomes necessary in learning a high number of parameters. Data augmentation and max pooling layers in the deep networks help to handle global and local spatial variations, which are mainly due to residual errors in head pose alignment, variations in physiognomy.

A second difficulty in practice is that of determining a suitable deep architecture. This effort must make various design choices, such as number of layers, type of units at each layer, layer size, number of convolution channels, etc. All these require considerable amount of experimentation, and training of each deep model requires very long periods and expensive computation resources (like servers with multiple GPUs) due to the large-scale data set sizes.

This review of the literature shows that, by means of transfer learning, deep models are applicable on relatively small labeled expression datasets. Nevertheless, they still need extremely-big training datasets for the learning of the initial deep model. The collection of 3D face datasets are much more difficult than simply collecting images from Internet. Although deep learning has recently been applied on 3D shape surfaces for object classification (Sinha et al., 2016), the resulting networks are quite shallow compared to the state-of-the-art in 2D images due to the limited sample-size. Because of this limitation, to the best of our knowledge, currently there is no work on deep models for 3D facial expression recognition. We stress that when training data is not abundant, the performance of deep architectures degrades significantly. For example, classification errors for object recognition on the Caltech-101 dataset using 5-layer Convnet (Zeiler and Fergus, 2014) without pre-training can be as low as 46.5%, while with pre-training on the on the ImageNet (Krizhevsky et al., 2012) with a set of 1.2 million images, the performance jumps up to 86.5%. However, in future, we expect to see competitive results with deep learning when transfer learning techniques are applied if sufficiently large datasets can be collected.

2.3. Face model-free vs. model-driven recognition

Expression recognition free of any face model avoids the complications arising from model-driven methods. First of all, there is no commitment to a prior face model, which itself could constitute a source of bias. For instance, the committed model, though conveniently fitted to the data, may not allow the use all discriminatory information on the expression face. Statistical shape models can be biased and restrictive depending on the chosen set of training faces. Deformation functions employed by the statistical models, such as the commonly utilized linear deformations, are often inadequate representations of many types of facial actions which are complex, local and subtle. Fully automatic model fitting can be error prone, and the quality of fitting directly influences the performance of the subsequent steps. Finally, the preparation of the models requires substantial and tedious manual work, expertise, and even additional datasets to be able to construct more generalizable statistical models.

On the positive side, model-driven recognition provides a crucial benefit which is not available to current face model-free methods. This essential advantage is the non-rigid registration of the face that reduces the confounding variations due to the misalignments of the facial/expression structures. However, even if the input face is aligned in pose, the facial/expression structures can still remain misaligned due to the physiognomy of individuals. Moreover, uncertainties and imperfections in pose alignment, especially in the presence of strong facial expressions, can augment residual pose alignment errors. As factors undermining model-driven methods, one can list non-linear transformations due to perspective projection onto 2D images, out-of-plane head rotations, and camera distortions. If the constructed models are adequate and the model-fitting works perfectly, model-driven methods can compensate for these confounding misalignments with varying degree of success, for instance, to a coarse level with the landmark-based methods, or to a more detailed level by using 3DMMs. In contrast, face model-free recognition methods provide some invariance to small local transformations while performing dense local patch-based analysis, since it becomes infeasible to learn all sorts of variations in the learning stage (Sariyanidi et al., 2015). Notice that, local patch-based analysis is also commonly used in model-driven methods to compensate for any imperfection in modeling and fitting stages, albeit it is used at sparse locations as guided by the face model instead of densely on whole image.

Recently, non-rigid registration has been applied in face model-free recognition by registering all the faces onto a common average luminance face image (Yang and Bhanu, 2012). However, this type of registration can only provide relatively coarse correspondence estimation due to the use of averaged (blurred) reference face, and it is used as an intermediate step for non-rigid face normalization before the extraction of conventional local patch-based feature extraction (LPQ descriptors). We show in the sequel that, via dense non-rigid registration on 3D surface geometry, pure data-driven facial expression recognition, i.e., without any face model and feature model, is possible. This is the key improvement over conventional face model-free recognition.

We emphasize that our work does not contest the superiority of model-free deformation estimation-based recognition over model-driven methods regarding to the recognition performance. We consider expression recognition using face models as a separate research field; instead our claim is that significant information can

be gained on the face deformation field leading to good expression recognition ability without requiring supervision of face models.

2.4. Dense shape model-free registration

Though face model-free and dense non-rigid registration have not been explored in facial recognition applications, shape model-free registration has actually been extensively studied and it is a maturing field, mostly in medical imaging for investigation of anatomical and temporal structures, for statistical population modeling, or for multi-modality fusion of different imaging devices or protocols. As discussed in the comprehensive taxonomy in Sotiras et al. (2013), there are three main aspects to be decided for any non-rigid registration method: the deformation model, matching criteria, and the optimization method. The most critical aspect is the choice of the deformation model, as it is a way to impose the priors by determining acceptable class of transformations, and as it allows to decide for the trade-off between computational efficiency and detail of description. The choice of matching criteria and of the optimization method is usually rather straightforward, though often depending on the task.

The deformation model acts as a regularizer of the ill-posed non-rigid registration problem. In the literature, various approaches ranging from knowledge-based methods to interpolation theory, from physical models to task-specific constraints have been proposed (Sotiras et al., 2013). The choice of deformation model becomes critical especially when the degree of ill-posedness is high and there are missing correspondences in the data. For face model-free recognition, we cannot utilize knowledge-based methods and task-specific constraints since they require use of face models. On the other hand, physical deformation models, like elastic deformations, can be applied as they do not require any face modeling. However the complexity and computational effort of physical models can be prohibitive, as even a single registration can take several minutes on a computer. This is especially due to the high number of parameters to infer. On the other hand, deformations derived from interpolation theory (radial basis functions, elastic body splines, free-form deformations, basis functions, and piecewise affine models) can be used to handle the transformations with much lower degrees of freedom, thus with lower level of complexity.

Registration of facial expression images without the guidance of face models may suffer from missing correspondences, i.e., when one of the images does not possess expression structure on one or more location, but which are present on the other image (e.g., furrows on the cheeks due to smile or disgust), and consequently the mapping (deformation) is not reliable at those locations. A straightforward remedy for shape model-free methods is to impose strong deformation priors (Sotiras et al., 2013) to somehow predict a local solution, however, at the cost of, possibly, detail loss and computational complexity.

In our work, our goal is not to register faces onto some common reference; instead, our goal is to detect the presence of local structures arising from expression deformations. Therefore we propose an alternative approach which employs multiple references. Thus, we find an optimal set of references guided by facial expression discrimination (via boosting). The underlying idea is that, while a single face reference can always suffer from missing correspondence cases (occlusions), the chance of finding local correspondences will increase with the increasing number of judiciously selected references. Once the optimal reference set is established, we then proceed to select the most discriminative pixels from different reference domains depending on the target *expression* to construct the classification features. Our features are, in fact, curvature values resampled by inverse deformation as detailed in Section 5. The multiple-reference scheme enables the use of simple and fast shape model-free registration for superior recognition. For highly detailed face model-free registration as well as fast processing we perform curvature intensity matching and multi-resolution gradient based continuous optimization as proposed in Savran and Sankur (2008). However, notice that our novel framework is generic regarding to the non-rigid registration routine, i.e., one can replace it with a shape model-free method of own choice.

3. Databases

We use three databases which differ not only in 3D image quality, but also they provide test-beds for different emotion/expression recognition scenarios.

The BU3DFE database (Yin et al., 2006) involves 100 subjects enacting the six universal emotions (happiness, sadness, anger, disgust, fear and surprise) (Ekman and Friesen, 1971). Four snapshots of each expression are taken during the subject's acting, where the first snapshot is the onset, the last one is the apex and other two are the in-between snapshots. We have included the apex and its preceding snapshot in our experimentation. Thus the total sample size is $100 \times (2 \times 6emotions + 1neutral) = 1300$. The faces have been captured by a high fidelity 3D structured light system. After segmentation, smoothing and sub-sampling, there are on the average 10K vertex points representing the face (Fig. 1).

The BOSPHORUS database (Savran et al., 2013;Bosphorus) contains images of 105 subjects which are labeled by a certified Facial Action Coding System (FACS) (Ekman et al., 2002) coder. The number of expressions per subject varies between 10 and 35, and these images involve both the universal expressions and instances of action unit (AU) combinations with different intensities. In this paper we consider detection of 25 AUs (18 lower AUs and seven upper facial AUs) selected from AU types that furnish sufficient sample size for experimentation. The total number of faces in our experiments is 2902. The faces have been captured by a high fidelity 3D structured light system. After segmentation, smoothing and sub-sampling, there are average 10K points representing the face (Fig. 1).

SBIA database (Savran et al., 2013) is comprised of emotional valence examples from 20 subjects, involving various positive emotions, such as joy, happiness, affection, pleasure and pleasant surprise, and anger, disgust, dislike, fear, startled surprise, and unpleasant surprise as negative emotions. It is a semi-spontaneous database and subjects are not constrained to frontal pose acquisition, unlike BOSPHORUS and BU3DFE databases. We use the apex frames in the experiments. The sample size is 707, which breaks down as 317 positive, 337 negative and 53 neutral VALENCE samples.

Unlike most of the prior 3D databases, SBIA database is acquired using a consumer-grade depth camera: The Kinect sensor. Kinect sensor provides inexpensive acquisition and it is simple in operation (small physical size and no need to train the operator). On the downside, such consumer-grade depth cameras present a big challenge, since their 3D quality is rather poor, i.e., high noise content and low resolution. Therefore, this database provides evaluation of the recognition performance on poorly captured images. After segmentation, smoothing and sub-sampling, there are on the average 3K points representing the faces (Fig. 1).

4. Fast non-rigid facial surface registration in 2D image domain

In order to simplify the computations and minimize the effect of resolution changes in 3D due to perspective effects, instead of directly deforming 3D faces, we first map 3D surface curvature onto a regularly sampled 2D image. Then we perform curvature intensity matching via gradient descent-based multi-resolution optimization.

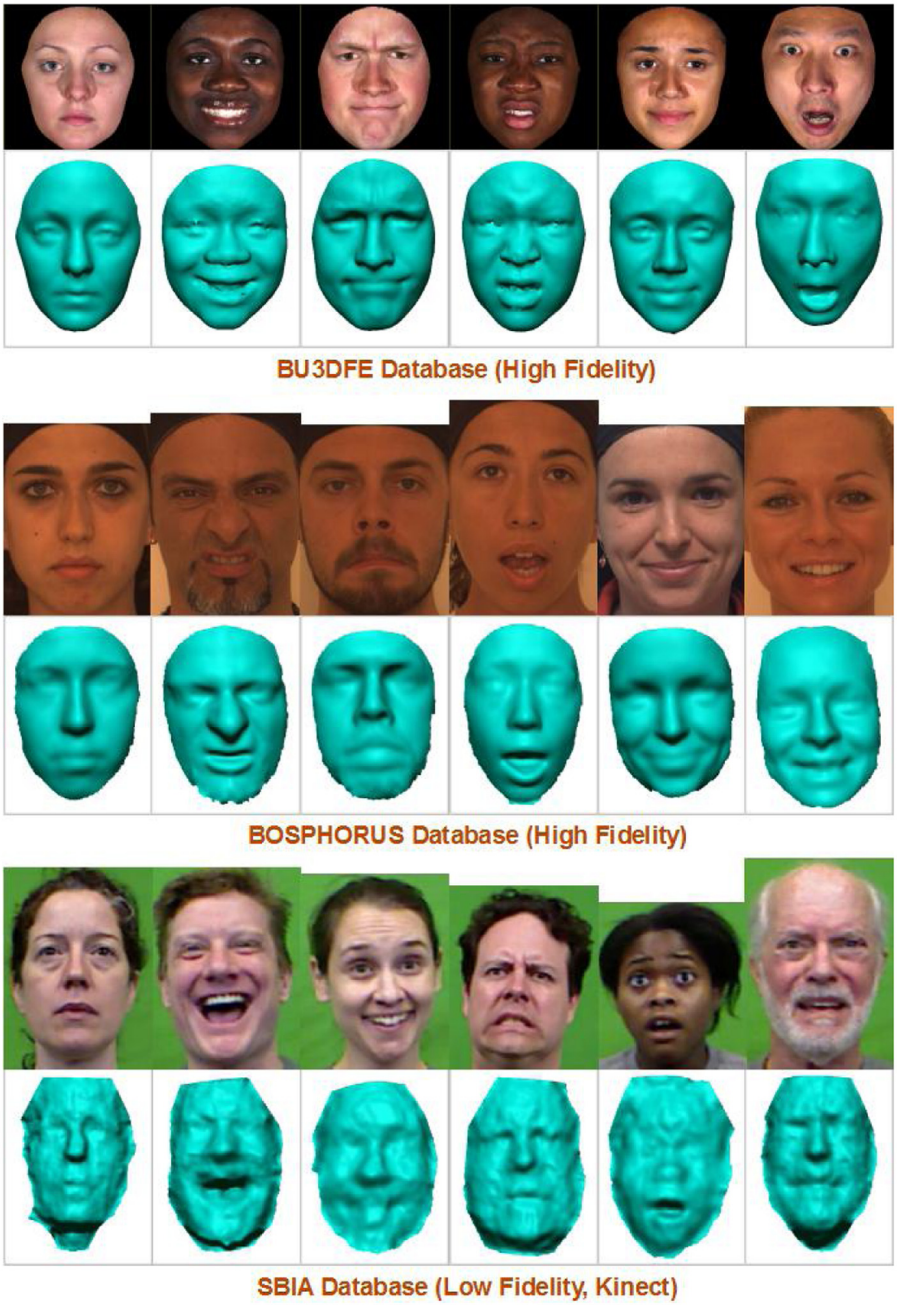


Fig. 1. Samples from the three databases.

4.1. Surface curvature image

The traditional local feature models used for recognition require spatial support to provide some degrees of translation invariance. The trade-off to achieve this invariance is potential information loss. Moreover, every feature model has different design parameters to choose, which may introduce bias. On the other hand, non-rigid registration estimates the local transformations, and thus it overcomes the requirement of translation invariance and allows purely data-driven analysis (i.e., feature-model free). Therefore we only consider features with minimal local context for our framework, i.e., primitive local surface quantities.

We can classify the primitive features according to the order of the differential operator to calculate them. For instance, depth or 3D coordinates are simply 0th-order; surface normals are 1st-order, and principal curvature based quantities - mean curvature, Gaussian curvature, shape index and curvedness - are 2nd-order primitive features. We prefer 2nd-order quantities since they provide direct and compact quantification of the local shape. This also means they are not affected by small residual pose variations due to imperfections of pose alignments, as they are invariant to rigid body motion. Being the trace of the shape operator (Gray, 1997), the scalar mean curvature compactly quantifies the bending of the normal field, hence the amount of local facial deformations. Our previous work has shown that mean curvature is the best 2nd-order feature among the four alternatives in all of AU detection, low-intensity subtler AU detection (Savran et al., 2012a), and AU intensity estimation (Savran et al., 2012b) problems. Therefore we employ in this work the mean curvature.

We generate surface curvature images after applying smoothing for the BOSPHORUS and SBIA databases; no filtering need to be done for the BU3DFE database since it has already been smoothed. Smoothing involves spike filtering via depth threshold, hole filling via morphological operations, and Gaussian smoothing on depth maps. We apply heavier smoothing on SBIA due to higher amount of noise. On BU3DFE and BOSPHORUS databases, we estimate curvature via mesh-based discrete estimation. However, due to the high amount of 3D noise in the SBIA database, we use a robust method by solving through normal curvatures as described in Savran et al. (2013).

The 2D curvature images are generated by re-sampling on a regular grid over the frontal faces, which have all been beforehand 3D pose aligned. We use the Iterative Closest Point (ICP) algorithm (Rusu and Cousins, 2011) to align faces onto a neutral face common to all subjects. Notice that pose alignment is an operation that is independent from feature extraction in our algorithm (which is a standard approach in facial expression recognition), and a system developer is free to choose his/her preferred method. Surfaces are represented as triangular wire-frames and we orthographically project the wire-frame meshes onto image planes after their alignment. The curvature values are re-sampled on the discrete image coordinates by calculating the triangular barycentric coordinates of the pixels and then by interpolating the curvature values at the triangle vertices. We have found that 96×96 pixel image resolution was adequate for subsequent processing. To find correspondences between grid pixels and mesh triangles with whom they are associated in a computationally efficient way for high density meshes, we render the meshes on the z-buffers of graphics hardware. After the re-sampling, regions of the grid that remain outside the 2D projection of mesh surfaces are filled by means of extrapolation (Savran et al., 2012a) to prevent abrupt changes at the domain boundaries.

4.2. Image matching by triangular discretization

In order to obtain the deformation between a pair of faces (their curvature images), we find a mapping, $\boldsymbol{\varphi}(\mathbf{p})$, satisfying image matching constraint between a reference image, I_{ref}, and a target image, I_{trg},

$$I_{ref}(\mathbf{p}) = I_{trg}(\boldsymbol{\varphi}(\mathbf{p})) \quad , \quad \mathbf{p} \in D_{ref}, \tag{1}$$

where $D_{ref} \subset \mathbb{R} \times \mathbb{R}$ is the 2D projection domain of the corresponding reference surface. Under the assumption that there is a full correspondence (bijection) between the pair of images, a mapping that satisfies the image matching constraint can be found by minimizing the image matching energy term

$$E_M(\boldsymbol{\varphi}) = \frac{1}{2} \int_{\mathbf{p} \in D_{ref}} \left(I_{trg}(\boldsymbol{\varphi}(\mathbf{p})) - I_{ref}(\mathbf{p}) \right)^2 \mathbf{dp}. \tag{2}$$

We minimize this energy functional for our registration purpose. However, this is a non-trivial problem. First of all, this is an ill-posed inverse problem, because we seek a 2D vector for each coordinate. Also, having similar curvature values at local regions, aperture problem can cause considerable errors. Moreover, there can be significant effects of the noise on the curvature values. As we explained in Section 4.1, we suppress the noise to a great extent via smoothing before curvature estimation which is adjusted at the calibration stage to match the requirements of the 3D sensor. Because of all these problems, it is impossible to estimate correspondences completely without errors. However, the error minimization strategy (coarse-to-fine gradient descent), the employed spatial constraint for the deformation, and the proper utilization of the estimated correspondences in the recognition algorithm, as to be explained in the sequel, all reduce the residual errors to an insignificant level for the recognition task. We show in Section 5.1 that a big information gain is achieved on various type of facial actions with this registration technique.

Our non-rigid registration performs locally affine interpolation using triangular meshes, which provides both a simple and fast solution as well as attains local spatial constraint for the deformation, i.e., local regularization. More explicitly, since the deformation gradient is constant over a triangle, moving a mesh vertex induces an affine transformation over each triangle that is connected to that vertex, thus provides a local spatial constraint for the deformation. We refer Sotiras et al. (2013) for a detailed discussion on the deformation models including the locally affine models.

Image matching energy (Eq. (2)) can be evaluated over a mesh of triangular elements as

$$E_M(\boldsymbol{\varphi}) = \frac{1}{2} \sum_{t \in T} \int_{\mathbf{p} \in D_t} \| I_{trg}(\boldsymbol{\varphi}_t(\mathbf{p})) - I_{ref}(\mathbf{p}) \|^2 \mathbf{dp} \tag{3}$$

where T is the set of all triangles, D_t is the domain of triangle t with the mapping function $\boldsymbol{\varphi}_t(\mathbf{p})$. The motion of triangle vertices ($\mathbf{p_k} \rightarrow \mathbf{q_k}$) implies locally affine motion, as evaluated by barycentric interpolation ($\widehat{(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)}$ denotes the triangle with vertices $\{\mathbf{p}_i\}$)

$$\mathbf{q} = \boldsymbol{\varphi}_t(\mathbf{p}) = \sum_{k=1}^{3} b_k(\mathbf{p}) \mathbf{q_k} \tag{4}$$

$$b_k(\mathbf{p}) = \frac{Area(\mathbf{p}, \widehat{\mathbf{p}_{i \neq k}, \mathbf{p}_{j \neq k, j \neq i}})}{Area(\widehat{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3})}. \tag{5}$$

Energy minimization is carried out by gradient descent with respect to mapping functional which is defined only on the mesh nodes. Let's define $\mathbf{q_n} = \boldsymbol{\varphi}(\mathbf{p_n})$ as the mapping on the mesh node n. Then the gradient at vertex n is obtained through the chain rule as

$$\frac{\partial E_M}{\partial \mathbf{q_n}} = \sum_{t \in T_n} \int_{D_t} b_{k(t,n)}(\mathbf{p}) \left(\frac{\partial I_{trg}(\mathbf{q})}{\partial \mathbf{q}} \Big|_{\boldsymbol{\varphi}_t(\mathbf{p})} \right)^T e_t(\mathbf{p}) \, \mathbf{dp}$$
$$e_t(\mathbf{p}) = I_{trg}(\boldsymbol{\varphi}_t(\mathbf{p})) - I_{ref}(\mathbf{p}). \tag{6}$$

Here, T_n is the set of triangles connected to the node n, $k(t, n)$ is the kth vertex of the triangle t that corresponds to node n, and $b_{k(t,n)}(\mathbf{p})$ is thus the kth barycentric coordinate (Eq. (5)) for the point $\mathbf{p}$. At every iteration of minimization, the mesh node n of the reference image is moved by the gradient vector, $\partial E_M / \partial \mathbf{q_n}$, which is estimated over the surrounding triangles $\{t \in T_n\}$.

The integrals in Eqs. (3) and (6) are approximated by re-sampling at the recursively subdivided triangle centers. This re-sampling procedure, however, is adapted to the area of triangles since mesh triangles can differ largely in area, by not allowing further subdivision if the area of the subdivided triangle is less than one pixel. Thus, while avoiding unnecessary computations for small triangles, we can accurately approximate integrals over the larger triangles. Bilinear interpolation is used for resampling from 96×96 image grids. To estimate the partial derivative terms inside the integral in Eq. (6), we use 3×3 Scharr masks. Since the degree of freedom of the deformation is constrained by the small number of mesh nodes and since we re-sample over triangular elements, this registration works quite fast.

Minimization of Eq. (2) is obtained through cascaded minimizations by starting from a coarse scale and then refining the solution at the consecutive finer scales, to avoid local minima and for faster convergence. This is realized by creating Gaussian image pyramids of the reference and target images, and by adapting the resolution of the reference domain meshes at each scale, as shown in Fig. 2 (see Savran and Sankur, 2008 for the details).

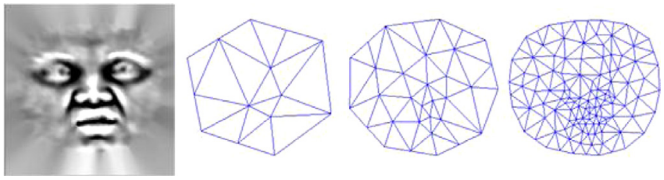


Fig. 2. Meshes of a reference curvature image adaptively generated according to three scales.

5. Non-rigid registration based data-driven recognition

We conjecture that detailed registration improves the information gain by reducing non-expression related variations and by increasing the discrepancy between the class (*expression*) conditional means, as will be explained soon in (9). We examine our conjecture by formulating the mutual information in Section 5.1. In the light of this conjecture, we develop a method which registers test images onto multiple references, as depicted in Fig. 3. Each registration is performed by deforming one reference toward the test image and then resampling the test image over the undeformed reference domain via inverse mapping. The boosted-references based recognition method is detailed in Section 5.2.

5.1. Improving information gain via registration

The goal of registration, as formulated in Eq. (2), is to minimize the sum of the squared differences between a reference and a target image. If all the images were deformed versions of each other with some small additive noise, the expected value of the square of difference (error) image over registered images $\mathbf{I}_r = \mathbf{I}_{trg} \circ \boldsymbol{\varphi}$, would have to be very small $E_{\mathbf{I}_r}[(\mathbf{I}_r - \mathbf{I}_{ref})^2] \approx \mathbf{0}$. This implies that the mean and variance of the registered images would be

$$\boldsymbol{\mu} = E_{\mathbf{I}_r}[\mathbf{I}_r] \approx \mathbf{I}_{ref}. \tag{7}$$

$$\boldsymbol{\sigma}^2 = E_{\mathbf{I}_r}[(\mathbf{I}_r - \boldsymbol{\mu})^2] \approx \mathbf{0}. \tag{8}$$

This assumption is valid if the differences between $\mathbf{I}_{trg}$ and $\mathbf{I}_{ref}$ can be removed via the operator $\boldsymbol{\varphi(p)}$ in Eq. (2). For instance, differences due to physiognomy can be estimated and removed to a great extent (see Fig. 4) since every face contains the same facial parts that play a role in strong correspondences, e.g., nose, eyes and mouth. However, under facial expressions and depending upon their strength, local violations of this assumption will occur. Our method is able to exploit both cases, that is, suppress physiognomic differences and profit from expression differences for classification. Therefore, we first examine their effects on the information gain.

For interpretation, we analyze the information gain computed for binary Normal Naïve Bayes classifier (NBC). We resort to NBC, first because, being a generative classifier, we can obtain analytical expression of the mutual information. The resulting numerical values provide simple interpretation and convenient visualization. Second, our *expression* recognizers are based on boosting of single feature NBCs.

Letting y and x denote the class label and scalar feature (e.g., curvature value of a selected pixel on a reference face) variables, respectively, the class conditional density is $p(x|y=c) = N(x|\mu_c, \sigma_c^2) = (2\pi\sigma_c)^{-1/2} \exp\left(-(x-\mu_c)^2/(2\sigma_c^2)\right)$. As we have derived in Appendix, the mutual information for binary NBC is approximated in terms of class means, variances and prior probabilities (p_0 and p_1) as

$$I(X;Y) \approx \frac{M}{V} = \frac{p_1 p_0 (\mu_1 - \mu_0)^2}{p_1\sigma_1^2 + p_0\sigma_0^2} \tag{9}$$

We now examine the information gain due to registration of *permanent structures* and *expression structures*.

Permanent face structures are always present on all the faces (with and without expressions, e.g., inner eye corners), that is, they are typical of all faces. Thus, the small variance assumption (Eq. (8)) holds at the pixels around those structures. Notice here that the term small is proportional to the magnitude of the structures in the expression images. Since *permanent face structures* are present in all classes, it follows then that their σ_0 and σ_1 are also small, which result in smaller variance V and higher information gain I (Eq. (9)); in other words, having compensated for physiognomic differences, the remaining differences of class means (M) are predominantly due to *expression* deformations. Fig. 5.a illustrates this effect clearly, that is, the role of registration in reducing the positive and negative group variances on faces. In the left part of Fig. 5.a, we see the mutual information (I) map for single pixel Normal Bayes classifier of AU24 - LipPresser according to the Eq. (9). The right part shows the same map after curvature images are registered. In the context of Fig. 5.a, the positive class denotes face images displaying AU24 with variations on the concomitant AUs; the negative class consists of all face images that do not contain AU24. As expected, variances estimated over non-registered samples are higher and more dispersed on the face than those estimated on registered samples, resulting in larger and darker regions. Also, the mean images of the registered samples look sharper. Not surprisingly, the strongest clues driving the registration of facial surfaces are the nose region and cavities around inner eye corners, which have correspondingly very small variance values. Other face regions provide less reliable clues due to the absence of strong structures or due to the variations originating from facial actions. As a case in point, mouth regions of the registered samples have almost always higher variance than any other parts of the faces.

Expression structures occur only with certain *expressions*, but are not otherwise present on neutral faces or on the remaining set of *expressions*. These structures play a marked role to drive the registration process provided they exist also in one or more of the reference faces bearing that *expression*. An illustrative case for *expression structures* case is the high curvature structure on the mouth that occurs during facial actions involving parting of the lips, as illustrated for AU16 - LowerLipDepressor in Fig. 6. Fig. 6 compares class conditional means (μ_1, μ_0) and the resulting mutual information maps (I) arising from two different references. The reference in the first row contains AU27 - MouthStretch, which shows a big opening of the mouth, and the second reference is that of a neutral face (lips are touching). We see from the class conditional mean images and their differences, M, that there are big differences between the class means around mouth for the AU27 reference (due to convergence of means, Eq. (7)), whereas this contrast is quite low for the neutral reference in the bottom row. We also see that both references yields low variance (V) maps since differences such as those due to physiognomy have been mitigated. The resulting I maps show that the reference with *AU27* provides much higher information gain.

In Fig. 5.b, we compare information gains due to non-rigid registration for various upper and lower face AUs. In general we see from the maps that registration increases the information gain by pushing informative (dark) regions into prominence and by attaining more marked values compared to the non-registered cases. These dark regions clearly reveal the AU related deformations on the face. Interestingly, AU27 - MouthStretch generates an unexpected map (rightmost bottom images in Fig. 5). While after registration, the information for AU27 lies correctly on the mouth region, without registration it is over all face parts. This is because mouth stretch is a very large deformation such that it affects the rigid registration of the face as vertical shift. Despite this global shift our registration yields a highly detectable pattern around the

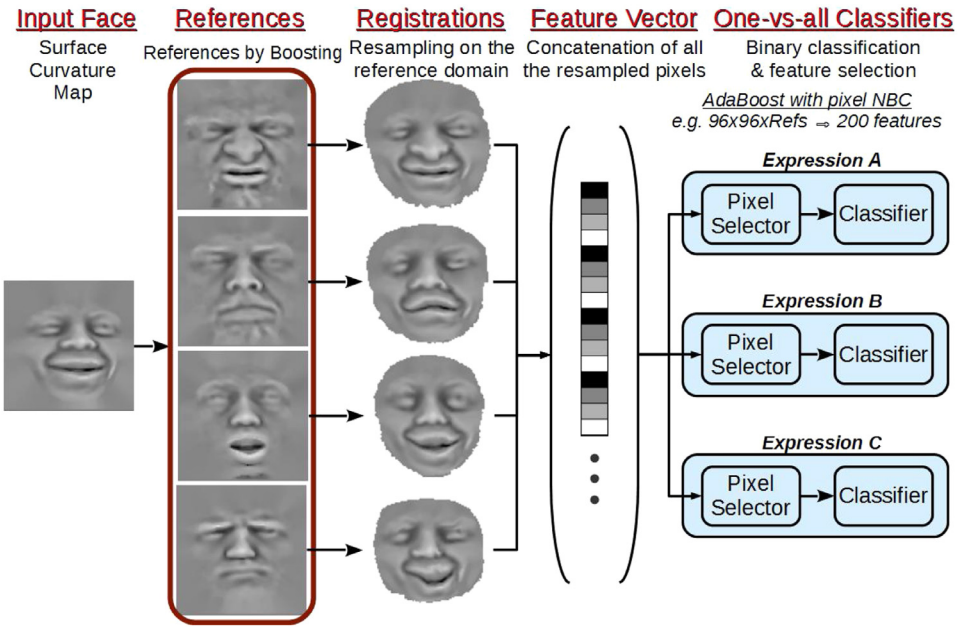


Fig. 3. Registration based data-driven recognition.

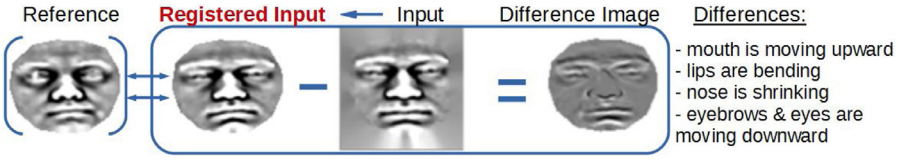
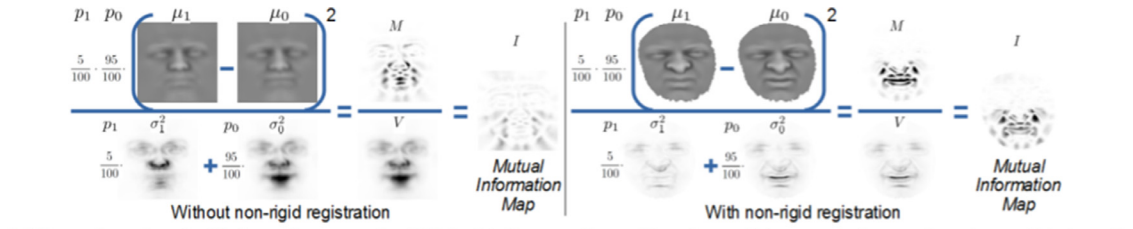
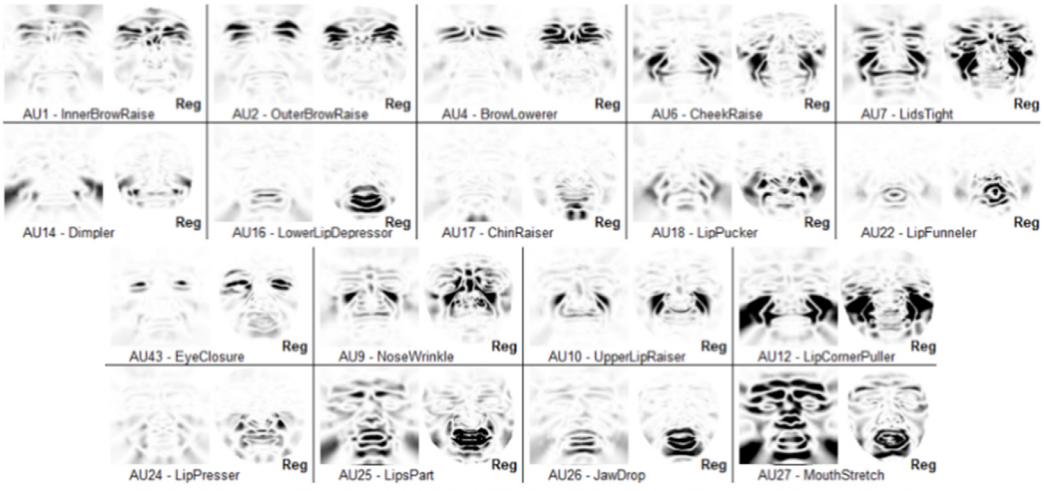


Fig. 4. Removal of physiognomy differences.



(a) Formation of mutual information maps for AU24 – LipPresser (1: positive class - AU present , 0: negative class - AU absent).



(b) Mutual information maps for different Action Units (Reg: with non-rigid registration).

Fig. 5. (a) Mutual information (I) map formation (Eq. (9)), and (b) action unit comparison examples on the BOSPHORUS database, for single pixel curvature Normal Bayes classifiers to evaluate non-rigid registration.

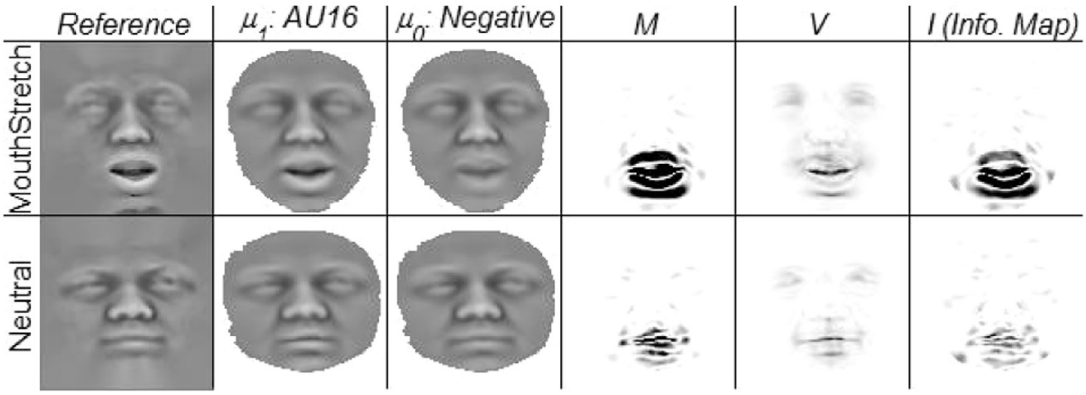


Fig. 6. Visualization of Eq. (9) using a neutral (bottom) and AU27-MouthStretch (top) references for AU16-LowerLipDepressor classification. Presence of open mouth in both AUs results in higher information gain (I map at top).

permanent facial structures. On the other hand, non-rigid registration compensates these misalignments and provides true location for the deformations.

5.2. Recognition via boosted references

As detailed in Section 5.1, there are two factors for improving expression detectability: low within-class variance via registration of *permanent face structures* and high between-class variance (differences of class means) via registration of *expression structures*. While the former goal is easily reachable with proper registration, the latter is non-trivial since *expression structures* cause missing correspondences as they are instance based. A common solution in shape model-free literature is to mitigate such errors by involving sophisticated deformation models (see Section 2.4). Instead of compensating for unavailable correspondences via strong deformation models, we approach the problem by incorporating multiple references, which can in total include a richer repertoire of local structures, alleviating thus the problem of correspondences. As shown in Fig. 5, a different reference can dramatically increase the information gain. We also show the improved information gain maps due to registration on optimally selected references for various expressions in Fig. 6. Therefore a proper selection of references is the key to the success of our approach. Fig. 3 depicts the overall method. An input face is registered onto a set of references, yielding pixel lists, that is, curvature values resampled by inverse deformation. Then, assigning NBC on each of these designated pixels, feature selection (e.g., 200 pixels) and classification are performed via AdaBoost.

The optimal selection of references is realized by a boosting based algorithm (Fig. 7). Since there has to be a common set of reference faces for all one-vs-all classification tasks (totally K tasks), at each iteration, we keep track of the boosting distribution of each task, B^k, but pick only one task to select a reference. The classification tasks used in the selection of the references are picked proportional to their positive sample sizes, i.e., according to the distribution $p(k)$ in Fig. 7. In principle this can also be achieved by randomly sampling the tasks from $p(k)$. However, when we sample only a few tasks (since we want at most like 25 references), correct proportional representation of the tasks in the reference set is not guaranteed. Therefore, we perform deterministic sampling using the scaled-histogram $s(k) = T_R \cdot p(k),\ k = 1, \ldots, K$ and give priority to classification tasks with larger positive samples. This is realized at every iteration by picking the mode of the current histogram, t'th task, and then updating the scaled histogram $s(k)$, for the next round of reference selection by decrementing the bin size of the selected task by one ($s(t) \leftarrow s(t) - 1$).

Having determined the task t at a boosting iteration, the best reference, r, is the one that contains the minimum error weak classifier for the task t under distribution B^t, and is added into the set R. Iterations are repeated till we have T_R references.

We start by considering all the training faces as reference candidates. This requires N^2 pairwise registrations. All the pairwise registrations are done before starting the boosting process. The storage of all the registered images in memory is not practical for the size of the databases considered. However, it suffices to store only the low dimensional estimated deformation parameters in the memory and execute the re-sampling whenever required during boosting. The training effort of the boosting can be expressed as $T_R \cdot D_a \cdot (N^2 + K)$ where D_a is the average domain size (number of domain pixel points). In our experiments, pairwise registrations of 707×707 pairs on SBIA, 1300×1300 pairs on BU3DFE, and 2902×2902 pairs on BOSPHORUS databases took 3.5 h, 12.5 h and 2.5 days, respectively on E5520 2.25 GHz workstation via parallel implementation on eight cores (including the duration for disk i/o). Thus, the resulting average pairwise registration is 0.025 s. On the other hand, the boosted reference selection algorithm on the SBIA, BU3DFE, and BOSPHORUS databases were completed in 14 min, 7.5 h and 11.5 h, respectively. In case the sample size is too high, training effort can be reduced, for instance by sub-sampling to reduce D_a or to reduce reference candidates to N_R so that we have $N_R \times N < N^2$ pairwise registrations.

6. Experiments

In order to facilitate experimental comparisons across different databases (Section 3) and classes with different priors, we calculate the area under the receiver operating curve (AuC) for each binary one-versus-all classification task; AuC is a threshold independent measure, invariant to prior probabilities of classes, and equivalent to theoretical maximum achievable correct binary classification rate. We evaluate hit rate (true-positives/all-positives ratio) versus false alarm rate (false-positives/all-negatives ratio). Tests are performed via 10-fold cross validation, where test subjects are not seen in the training sets and each fold is forced to be balanced with respect to positive sample sizes, since their numbers are much smaller than the number of negatives. Then for each *expression* mean and standard deviation of AuC's are estimated over the test folds. Finally, averaged AuC and standard deviations are calculated by weighted averaging where weights are proportional to positive sample sizes. We also report the correct classification rates when comparing with the former results in the literature.

In the description of experimental results, we first examine and discuss the benefit, if any, of the regularization model, applied

T_R : Number of references to be selected
R : Set of selected references
N : Face image sample size
K : Number of binary classification tasks
N_1^k and N_0^k : positive and negative sample sizes for task k
$\mathbf{I}_i : \mathbb{R}^{w\times h}$: Curvature image of the i^{th} face ($w \times h$ is image size)
D_i : Image domain of the i^{th} face
$S = \{y_i^k, \mathbf{I}_i, \boldsymbol{\varphi}_{ij}\ ;\ k = 1...K; i = 1...N; j = 1...N\}$
$y_i^k \in \{0, 1\}$: Class label of sample i for classification task k $\mathcal{I} = \{\mathbf{I}_i \in \mathbb{R}^{w\times h} \mid i = 1...N\}$ where $w \times h$ is image size
$\varphi_i = \{\boldsymbol{\varphi}_{ij} : D_i \to \mathbb{R} \times \mathbb{R} \mid j = 1...N\}$ where $\boldsymbol{\varphi}_{ij}$ is registration mapping from reference image i to target j and is defined over the reference domain D_i
$h^k(x) : \mathbb{R} \to \{0, 1\}$: Normal Bayes classifier (NBC) for task k
B^k : Boosting distribution of classification task k

procedure BOOSTEDREFERENCESEL(T_R, S)

< Initializations >
  for $k = 1 \cdots K$ ▷ for each binary classification task
    $p(k) = \frac{N_1^k}{\sum_l N_1^l}$ ▷ Probability of selecting task k
    $s(k) = T_R \cdot p(k)$ ▷ Scaled histogram for task k
    $B^k(i) = \begin{cases} 1/N_1^k, & y_i^k = 1 \\ 1/N_0^k, & y_i^k = 0 \end{cases}$
    Normalize B^k to be a distribution
  end for

< Iterate until T_R references are selected >
  for T_R iterations
    $t \leftarrow \arg\max_k \{s(k)\}$ ▷ Pick a classification task
    $s(t) \leftarrow s(t) - 1$ ▷ Reduce its histogram bin
    for $i = 1 \cdots N$ ▷ For each image in $\mathcal{I}$
      for each $\mathbf{p} \in D_i$
        $\varepsilon_k(i, \mathbf{p}) \leftarrow$ CALCERROR($\{y_i^t\}$, B^t, $\mathcal{I}$, φ_i, $\mathbf{p}$)
      end for
    end for
    $r \leftarrow \arg\min_i \left\{ \min_{\mathbf{p}} \{\varepsilon_t(i, \mathbf{p})\} \right\}$
    ▷ Index of the reference with the minimum error
    $R \leftarrow R \cup r$

  < Update distribution for each classification task >
    for $k = 1 \cdots K$
      for each $\mathbf{p} \in D_r$
        $\varepsilon_k(r, \mathbf{p}) \leftarrow$ CALCERROR($\{y_i^k\}$, B^k, $\mathcal{I}$, φ_r, $\mathbf{p}$)
      end for
      $\mathbf{p}^* \leftarrow \arg\min_{\mathbf{p}} \{\varepsilon_k(r, \mathbf{p})\}$
      $\varepsilon_k^* \leftarrow \varepsilon_k(r, \mathbf{p}^*)$
      $h_i^k \leftarrow h^k(\mathbf{I}_i \circ \boldsymbol{\varphi}_{ri}(\mathbf{p}^*))$
      $B^k(i) \leftarrow B^k(i) \cdot \begin{cases} (1 - \varepsilon_k^*)/\varepsilon_k^*, & y_i^k \neq h_i^k \\ 1, & y_i^k = h_i^k \end{cases}$
      Normalize B^k to be a distribution
    end for
  end for

  return R
end procedure

procedure CALCERROR($\{y_i\}$, B, $\mathcal{I}$, φ, $\mathbf{p}$)

  $P = \{\mathbf{I}_j \circ \boldsymbol{\varphi}_j(\mathbf{p})\}$: Set of resampled features at $\mathbf{p}$
  Train NBC over P with distribution B
  Calculate error ε over P with distribution B
  return ε
end procedure

Fig. 7. Boosted reference face selection algorithm.

at varying strengths, on the registration outcome (Section 6.1). Then, we evaluate our boosted reference selection algorithm (Section 6.2). In Section 6.3, we investigate various aspects of recognition by non-rigid registration based features, comparing against commonly used face model-free feature extraction. Finally, we compare the recognition performance of our method with the previous work, as given in Section 6.4, based on AuC and correct classification rates (Fig. 7).

6.1. Evaluation of registration

In order to validate the effectiveness of our registration method as well as to prove our conjecture, where we claim that strong global regularization is not required if such a multiple registration-based scheme is used (since it avoids critical missing correspondences), we perform a series of comparative evaluations. Being an inverse ill-posed problem, non-rigid registration requires regularization to constrain the solution and to obtain a smooth deforma-

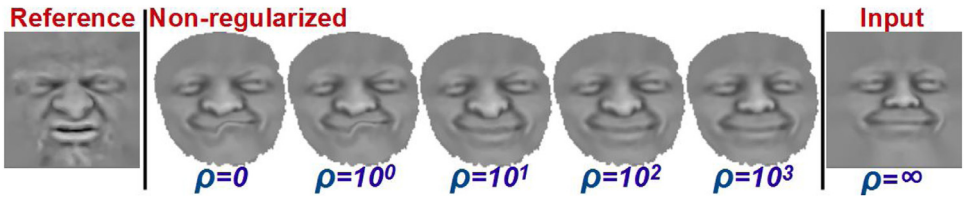


Fig. 8. Registrations for different rigidity values, ρ.

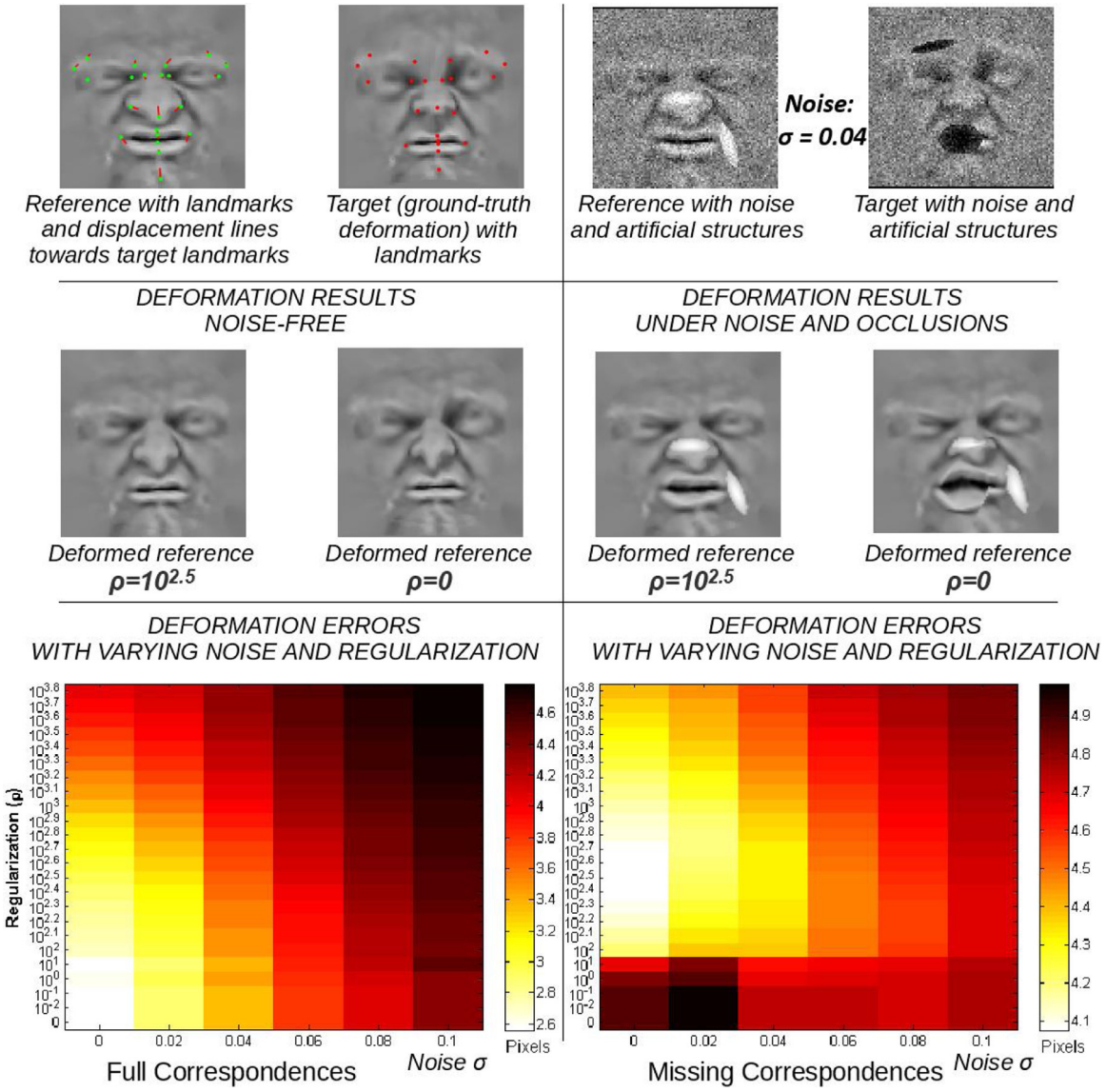


Fig. 9. *First row*: reference and target faces; target faces are generated by applying ground-truth deformation obtained with TPS interpolation, which maps reference landmarks (green) onto target landmarks (red). *Second row*: deformation of the reference with regularization $\rho = 10^{2.5}$ and $\rho = 0$. *Third row*: estimation error of deformations averaged over 90 registrations, all replicated with varying values of regularization parameter and additive Gaussian noise variance. *Left column*: Full correspondence case; *Right column*: missing correspondence case. (Details with various example deformation estimations are available in the supplementary document.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tion field (Sotiras et al., 2013). To test the effect of regularization in face registration, we augment the objective energy functional ($E_M(\boldsymbol{\varphi})$ in Eq. (2)) with a deformation energy functional, $E_D(\boldsymbol{\varphi})$, resulting in the total energy functional $E_T(\boldsymbol{\varphi})$,

$$E_T(\boldsymbol{\varphi}) = E_M(\boldsymbol{\varphi}) + E_D(\boldsymbol{\varphi}). \tag{10}$$

To handle large expression deformations, we use a hyper-elasticity model, which is better than standard linear elasticity, in the deformation energy functional. The potential energy is measured by the Green-Lagrange strain tensor

$$\mathbf{E}_{\mathbf{GL}}(\boldsymbol{\varphi}) = \frac{1}{2}\left(\frac{\partial \boldsymbol{\varphi}}{\partial \mathbf{p}}^{\mathbf{T}} \frac{\partial \boldsymbol{\varphi}}{\partial \mathbf{p}} - \mathbf{I}\right), \tag{11}$$

and we evaluate the deformation energy by its Froebenius norm with the weight coefficient ρ over triangular meshes (Savran, 2011)

$$E_D(\boldsymbol{\varphi}) = \rho \int_{\mathbf{p}\in D_{ref}} \|\mathbf{E}_{\mathbf{GL}}(\boldsymbol{\varphi}(\mathbf{p}))\|_F^2 \, \mathbf{dp}. \tag{12}$$

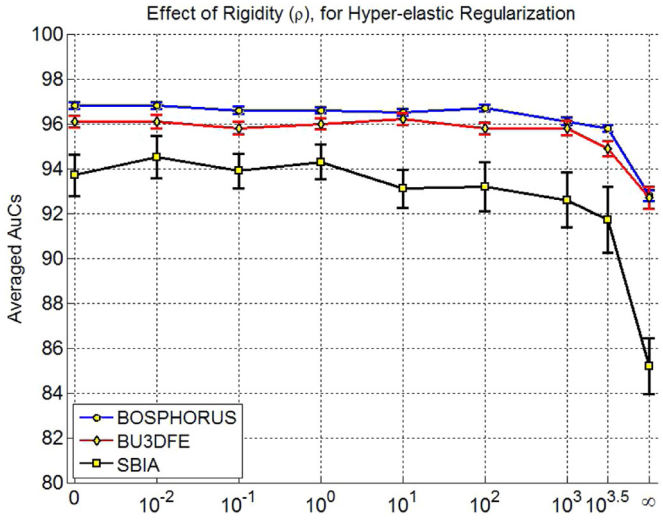


Fig. 10. Recognition performances (AuC with standard deviation bars) for varying ρ.

Deformation energy and its gradients are also computed on the triangular meshes (Savran, 2011). This regularization term can provide estimation of physically plausible face deformations by appropriate choice of ρ, since it establishes diffeomorphism, i.e., a smooth (C^∞) and bijective mapping with a smooth inverse. ρ controls the amount of rigidity. While bigger ρ values make the deformations more rigid, hence providing strong regularization and possibly avoiding non-plausible results, such registrations may not compensate for all local mismatches, especially if the actual deformations of the input face are large vis-à-vis the reference. On the other hand, while with small ρ the images can be matched better locally, non-plausible deformations may also occur due to the absence of corresponding structures. As can be seen in Fig. 9 we obtain an unrealistic deformation of lips since correspondences are missing between open mouth (reference) and a closed mouth in the input image (see also Figs. 3 and 11).

For the accuracy evaluation, we have prepared ground-truth deformations by applying thin-plate-spline (TPS) interpolation driven by 22 facial landmarks marked on 10 samples (of different subjects and expressions) from the *The BOSPHORUS database*. We have also added Gaussian noise as well as artificial structures to test the robustness and handling of missing correspondences. An example reference-target pair is shown in the first row of Fig. 9, with corresponding deformed references for two values of the rigidity parameter ($\rho = 10^{2.5}$ and $\rho = 0$) in the second row. Accuracy is evaluated by calculating the average Euclidean distance between estimated and true dense deformation fields (displacement vectors) on $10 \times (10-1) = 90$ registration pairs.[3] The third row in Fig. 9 shows the estimation errors of the full deformation fields in pixels under different regularization and noise levels, as well as for missing correspondence cases. We observe that accurate registration is obtained under noise without any need for global regularization, and that global regularization (larger ρ) is helpful only when there are missing correspondences. See for example the darker left bottom corner of the sub-figure Missing Correspondences in Fig. 9; the performance improves rapidly with the rigidity parameter increasing beyond $\rho > 10^0$.

Second, we calculate the average recognition performances, shown as a function of ρ (with the optimal references found in Section 6.2) in Fig. 10. We observe consistently across all three databases that, although the appearance of registered faces can differ depending on the chosen rigidity value (see Fig. 8), the recognition performances are insensitive to ρ, until $\rho = 10^3$ where registration becomes ineffective due to strong penalization of deformations. In fact, we obtain the best results for $\rho = 0$.

Both deformation accuracy and recognition performance evaluations reveal two important outcomes. First, they show that, without using global regularization, local affine interpolator, which inherently acts as local regularizer in our registration method, is sufficient to handle ill-posedness on facial curvature images. Moreover, even though the global regularization is necessary to mitigate errors due to the missing correspondences (Fig. 9), it is not necessary to improve recognition (Fig. 10). Thus we provide strong evidence for the validity of our conjecture that strong global regularization is not required and for the effectiveness of our non-rigid registration method.

6.2. Evaluation of boosted references

We evaluate the boosting-based reference selection by comparing against a simpler method, which basically picks references from dataset proportional to sample size of *expressions*. This is accomplished by using scaled-histogram $s(k)$ (in lines 9–10 in Fig. 7) to pick the binary classification task t. Then a reference is randomly selected from the positive samples of the task t. Hence, this method is called histogram-based reference selection method.

Fig. 12 shows the AuC performances for varying number of references. This figure is a clear evidence on the necessity of multiple-references, as use of single reference always results in lower recognition performance. We observe rising trends in performance with increasing number of references across databases. Boosted-based reference method is slightly superior to histogram-based method of selecting references.

We fixed the number of registration references according to this evaluation. There is a trade-off between recognition performance and computational efforts, both increasing with the number of references. Therefore, as a compromise, we pick 11, 9 and 13 references for the BOSPHORUS, BU3DFE and SBIA databases, respectively (see Fig. 11 for selected examples).

6.3. Evaluation of recognition

We compare our method with Gabor filter-banks and LBP descriptors, following common practices. For the Gabor wavelets, four orientation and five scales, corresponding to wavelengths from 4 to 16 pixels, are applied. Responses of those 20 wavelets are computed densely at each image pixel and their magnitudes are used as the features. For the LBP features, images are divided into non-overlapping blocks where binary pattern histograms (59 histogram bins by using only the uniform patterns) are computed. Best LBP radius and block size parameters can be quite different for different databases (Shan et al., 2009). Therefore, we determined these parameters on the three databases by experimenting over all the combinations of 6×6, 8×8 and 12×12 block divisions and radius sizes of $r \in \{1, 2, 3, 4, 5\}$ pixels. We found that the best parameters on the Bosphorus and BU3DFE parameters are the same, 12×12 block partitioning and radius $r = 4$, while the optimal parameters for the SBIA database are 8×8 partitioning with radius $r = 4$. Benefit of larger block size on the SBIA database may be due to the resulting smoothing effect on the highly noisy data from consumer depth cameras.

In AdaBoost based recognition on high dimensional inputs, the convention is to employ simple classifiers as weak learners, like tree-stumps or linear NBC with shared-variance. However, we observed that quadratic discrimination, i.e., NBC without shared-variance, considerably improves recognition as seen in Fig. 13, especially on curvature images (pixel features) that had

[3] See the supplementary document for the experiment details.

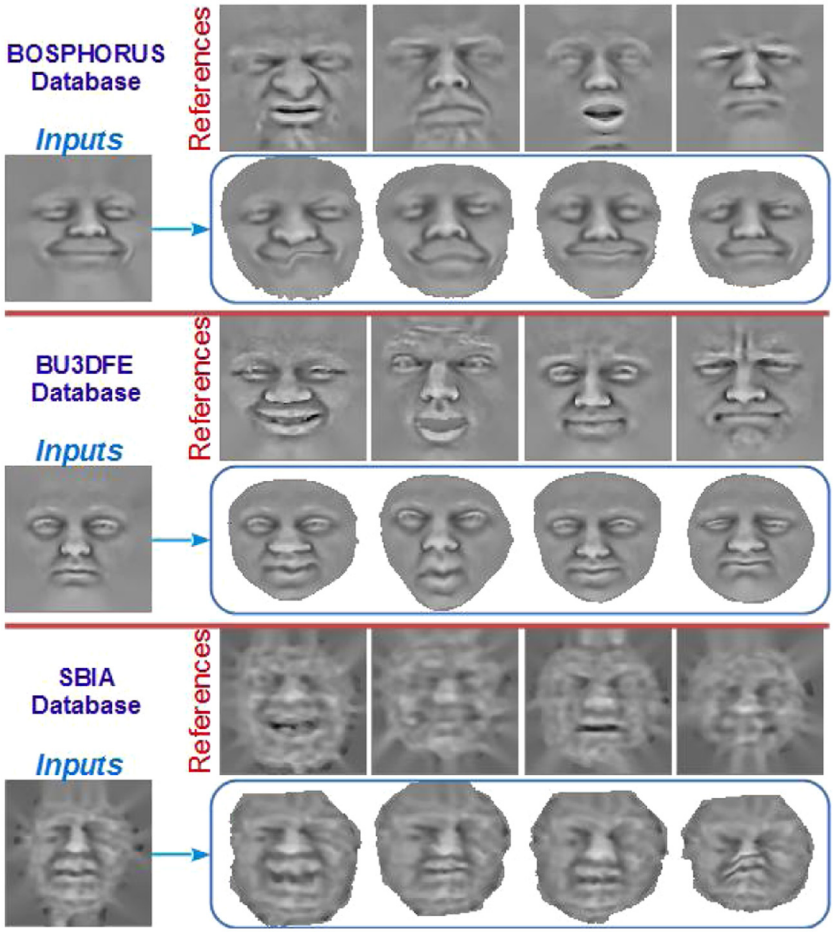


Fig. 11. Examples of registered curvature maps onto the first four references that are selected by boosting.

the lowest performance. We also observed that when quadratic NBC is used, AdaBoost achieves higher performance than state-of-the-art AdaSVM (SVM on the selected features by AdaBoost) (Littlewort et al., 2011; Sandbach et al., 2012a; Savran et al., 2012a). For instance, it achieved 96.8% AuC score on BOSPHORUS and 96.1% on BU3DFE with registration based recognition, whereas AdaSVM achieved slightly lower scores, 96.1% and 95.6%, respectively. Marginals improvement by increasing the number of selected features is possible as seen in Fig. 14 albeit with strongly diminishing returns (e.g., 97.2% AuC score on BOSPHORUS with registration based recognition). We fixed the number of features to 200 to run in feasible duration the large number of experiments.

When we look at the performances of all the features in Fig. 13, we see the superiority of non-rigid registration-based features. Second in performance ranking comes the Gabor filter-banks. Moreover, we see that the registration-based method does not deteriorate as much compared to others under adverse conditions, i.e., when a linear classification is used in lieu of a quadratic one, when applied on low-fidelity 3D data of consumer depth cameras (SBIA), or when the number of features is reduced (Fig. 14). As a final remark, comparing individual one-vs-all AuC performances in Fig. 15, we see that registration-based recognition performs better for almost all types of *expressions*.

Finally, we evaluate resiliency of our recognition method against residual errors of pose alignment. In fact, perturbations, even small, in pose is one of the major causes that degrades recognition performance, whether it is for luminance data or for 3D geometry data. This is because there is always some uncertainty in the face alignment, which is amplified in the presence of some expressions. In fact, existing face model-free local feature extraction methods try to compensate for pose perturbations by using large windows to provide some degrees of shift-invariance (Sariyanidi et al., 2015), which, however, causes loss of details. Our method does not need local windowing and each feature corresponds to a single pixel, thanks to the local shift-invariance provided by the detailed registration; hence we do not incur into any loss of detail. We test the conjecture that non-rigid registration-based method is robust to pose perturbations in an experiment where the performance with ICP alignment is compared with the performance resulting from a more accurate alignment using manually annotated facial landmarks. Such landmarks are available in BOSPHORUS and BU3DFE databases. Fig. 16 compares the recognition performances of the manual and of the automatic face alignments. While the performance of the non-rigid registration-based scheme remains the same under both types of alignment, all the other feature types perform worse in the ICP alignment (less accurate). Notice that there is practically nothing that our method can gain from manual landmarks, and hence more accurate registration since it is fully compensating for all pose discrepancies.

This compensation ability also eliminates any bias due to the arbitrary choice of an ICP alignment reference. Even though we have chosen a random neutral face as the alignment reference, our method allows the use of any expression face. Especially, expression faces with large deformations, like big mouth openings, can bias a rigid transformation. We can see the effect of this bias at the bottom-right corner of Fig. 5 which compares the information gain maps on the mouth stretch action (AU27) with and without non-rigid registration. We see that, unless dense non-rigid registra-

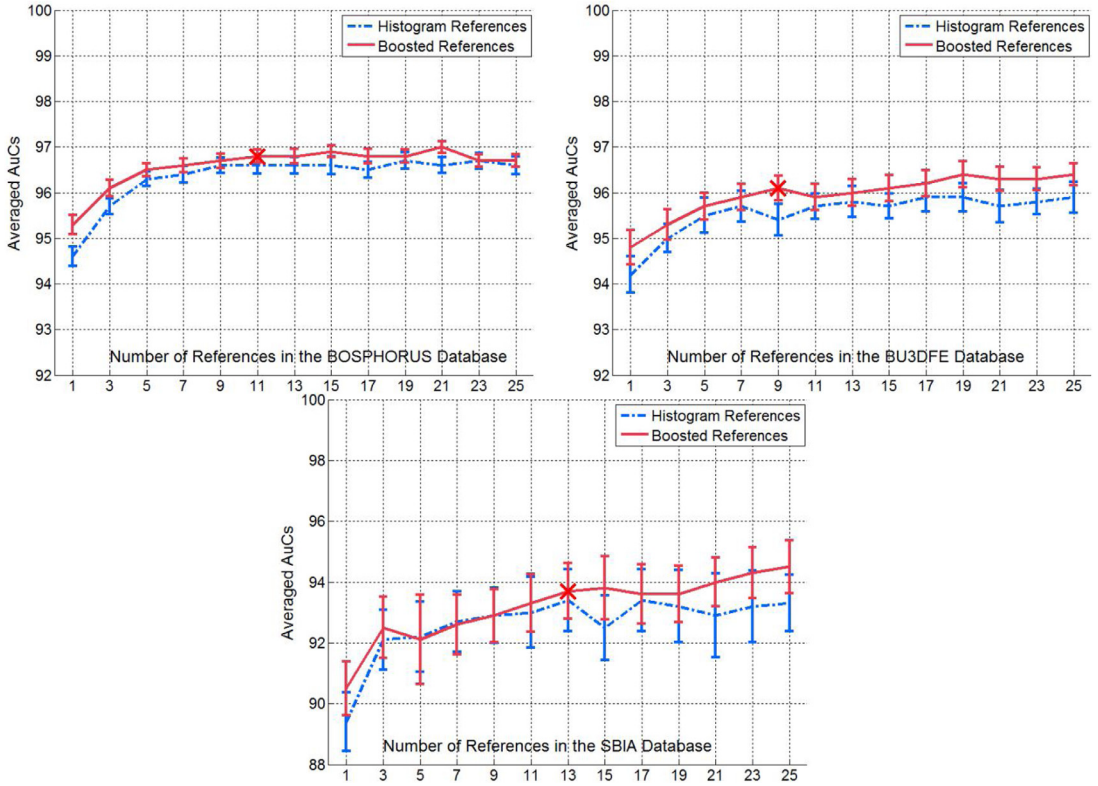


Fig. 12. Recognition performances (averaged AuC values ± standard deviation bars) of references selected by boosting and by histogram sampling. The chosen number of boosted references are marked by cross signs.

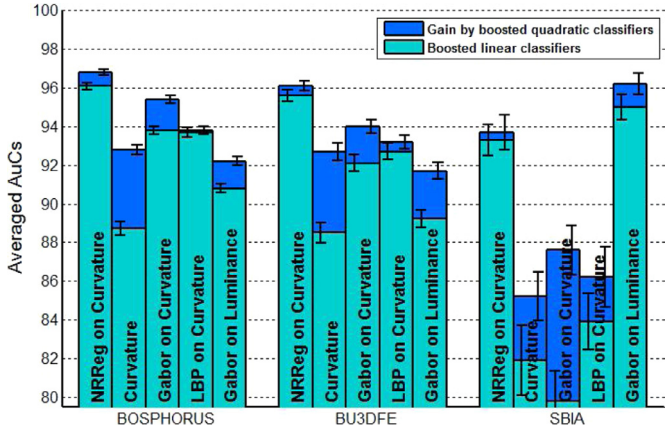


Fig. 13. Recognition performances (averaged AuC ± standard deviation) with different features, and with boosted linear and quadratic Bayes classifiers.

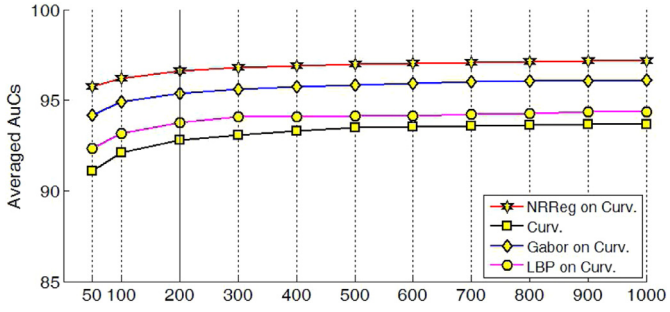


Fig. 14. Recognition performances (averaged AuC) under varying number of features (BOSPHORUS db.).

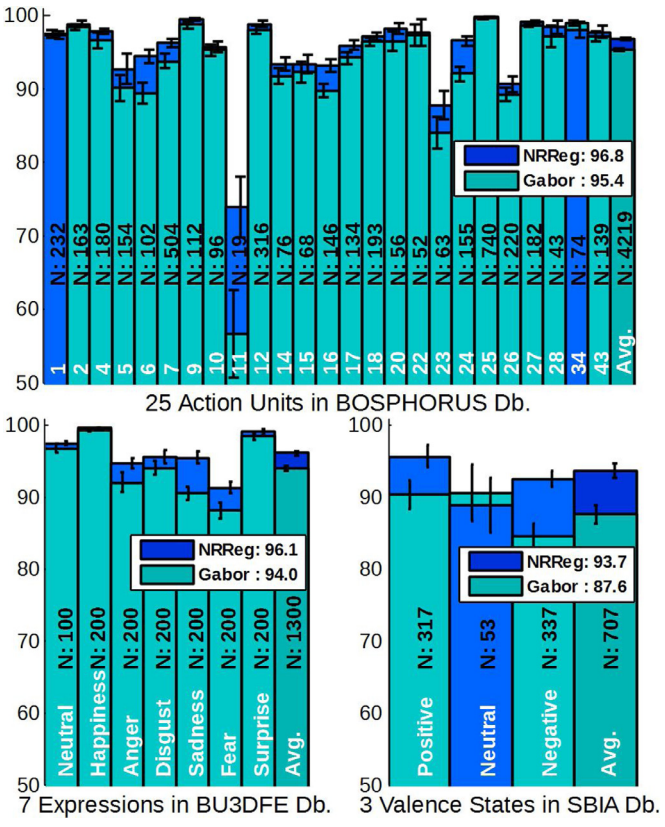


Fig. 15. One-vs-all AuC scores (with standard deviation bars) of non-rigid registration and Gabor features on surface curvature images (N: sample size).

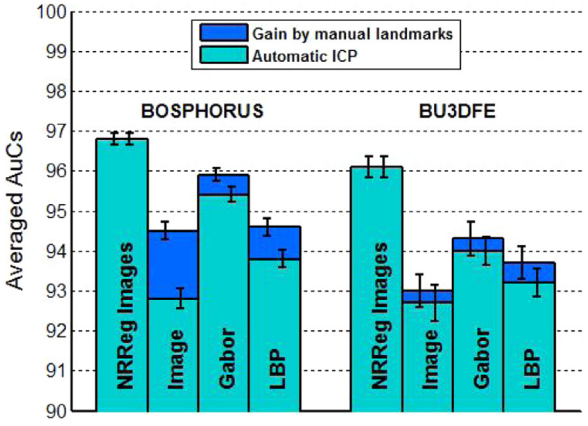


Fig. 16. Averaged AuC values (with std. bars) of recognition by manual landmark versus automatic ICP alignment for different feature extraction methods.

Table 1
Averaged [AuC - correct rate] results (Curv: curvature image, Lum.: luminance image, NR: Nonrigid).

Features (200)	BOSPHORUS	BU3DFE	SBIA
NR Registration	96.8–96.4	96.1–83.2	93.7–85.4
Curv.	92.8–93.5	92.7–73.2	85.2–76.0
Gabor on Curv.	95.4–95.7	94.0–76.0	87.6–77.2
LBP on Curv.	93.8–94.1	93.2–74.4	87.1–76.7
Gabor on Lum.	92.2–93.1	91.7–72.3	96.2–84.9

tion is utilized, the information becomes spread all over the face and loses its locality instead of concentrating correctly over the mouth region. This spread of the deformation information is because targets have large deformation differences with respect to employed neutral face alignment reference and these cause vertical shifts in all the mouth stretch samples. Since the non-rigid registration successfully compensates for this vertical shift, the information remains localized around the mouth region.

6.4. Comparison with the prior work

In order to make comparisons with the previous work, we also calculate the recognition rates and the confusion matrices. Table 1 shows the recognition rates and the AuC scores. We observe that all features follow the same trend under both metrics.

Previous AU detection performance on the BOSPHORUS database in terms of AuC was: 96.3% with Gabor wavelets on mean curvature and shape index data (Savran et al., 2012a), 96.3% AuC by utilizing LBPs on normal vectors (Sandbach et al., 2012c), and 97.2% by combining LBP and Gabors as local-depth-Gabor-binary-patterns (Sandbach et al., 2012a). However, Sandbach et al. (2012a); 2012c) had achieved these scores by manual landmark-based pose alignment. As we have shown in Section 6.3, Fig. 16, in reality, the residual alignment errors degrade the performances considerably. In contrast, with fully automatic alignment, thanks to its compensation capability for residual pose, non-rigid registration obtains 97.2% with 1000 features (Fig. 14); and if the feature count is limited to 200, it still achieves a high score of 96.8% AuC (Table 1).

There have been various reported results on the BU3DFE database. In chronological order the correct recognition rates are as follows: 83.6% by Wang et al. (2006), 91.3% by Soyel and Demirel (2007), 95.1% by Tang and Huang (2008), 77.5% by Berretti et al. (2010), 98.8% by Maalej et al. (2011), and 73.0% by Vretos et al. (2011). Notice that, except for Vretos et al. (2011), all the other methods owe their high recognition performance to feature extraction aided by manual landmarks (between 20 and 83 landmarks).

Table 2
Confusion matrix on the BU3DFE database using the NR Registration method.

	Neu	Hap	Ang	Dis	Sad	Fea	Sur
Neu	87.5	2.0	6.0	0.0	3.0	1.0	1.0
Hap	0.0	94.5	0.0	0.0	0.0	4.5	1.0
Ang	2.5	0.5	77.0	6.0	9.0	5.0	0.0
Dis	1.5	1.5	6.5	79.0	2.5	6.0	3.0
Sad	4.5	0.0	8.0	0.0	82.0	5.5	0.0
Fea	3.0	9.5	4.5	6.50	4.5	68.0	4.0
Sur	0.5	0.5	0.0	2.0	0.5	3.5	93.0

Table 3
Confusion matrix on the SBIA database using the NR Registration method.

	Pos (317)	Neu (53)	Neg (337)
Pos	90.9	0.0	9.1
Neu	11.3	40.1	48.6
Neg	11.6	0.6	87.8

However, as recently shown in a comprehensive study (Creusot et al., 2013), actually landmark detection on 3D faces is highly error prone, especially in the presence of the *expressions*. For instance, on the BOSPHORUS database, average detection rates can drop down to about 65% (even down to 25% for the chin point). Moreover, the detection rates considerably degrade under out-of-plane rotations of the face. Therefore, features based on automatically extracted landmarks are prone to performance drops under expressions and rotations. As a case in point, Maalej et al. (2011) show that when only moderate noise is added on the eyebrow landmarks, the performance drops from 98.8% to 85.6%. It is obvious that this drop will be more dramatic, under high localization errors, especially of the difficult lower face landmarks (Creusot et al., 2013). Our non-rigid registration-based feature extraction achieves 83.2% with automatic pose alignment and fully automatic feature extraction, surpassing its nearest competitor (Vretos et al., 2011), who achieve 73.0% correct rate via Zernike-moments. The confusion matrix in Table 2 shows that most of the errors happen with the expression *fear*.

On the SBIA database, prior work (Savran et al., 2013) has obtained correct rate of 77.4% using histogram-based descriptors on the mean curvature data; this score is considerably lower than our correct rate of 85.4%. Table 3 shows that *neutral* class is often confused with *negative* valence. This can be explained by its relatively small sample size (53 vs. 317 and 337).

7. Discussions

In this section we discuss several important points that can be useful when implementing an algorithm based on our generic deformable framework for face model-free recognition. These points are: choice of the non-rigid registration technique, an extra step of high-level feature extraction after registration, and extension to applications on spontaneous expressions and uncontrolled environments.

7.1. Choice of the non-rigid registration technique

As we discussed in detail in Section 2.4, a variety of shape model-free non-rigid registration methods has been proposed (Sotiras et al., 2013). One could want to investigate these alternative registration techniques as for their potential to alleviate the ill-posed registration problem and to mitigate the errors due to missing correspondences via the imposition of strong priors. However, their direct application for expression recognition may be precluded by the computational complexity of deformation models.

Instead, our proposed solution uses a simple registration method on a multitude of registration references. This can handle, to a large extent, the missing correspondence problem, as shown experimentally in Section 6.1. This flexibility enables us to run the registration without strong priors, and makes our method computationally feasible despite the dense registration approach. In other words, simple and fast registration methods, like the one we propose in Section 4, are adequate in our multiple-reference framework. Recall that we obtain improved recognition performances even though our registration may not yield case by case the best image-pair registration. This is in accordance with some works in the literature suggesting the use of simplest registration methods for the recognition tasks. After comparing various deformation models for different recognition problems, Keysers et al. (2007) have shown that the simplest image distortion model performs as good as the more constrained models, with the crucial advantages of low-computational complexity and simplicity of implementation. Consequently, based both on the literature and on our experimental evaluations, we suggest the use of the simplest registration methods if the recognition framework somehow involves a representative set of references.

7.2. High-level feature extraction after registration

As we have shown in Section 6.3 (e.g., Fig. 13), despite the fact that we have only used registered curvature pixels and no higher-level features, our method surpasses the high-level feature models. Recall that the latter encodes the local patterns of the curvature maps. Then a natural question follows: since high-level feature models considerably improves the performance on the undeformed curvature maps, can they also bring improvements if applied on the registered curvature pixels? The short answer according to our preliminary experiments (Savran, 2011) is negative. The result of our previous work have shown that dense Gabor feature extraction on each deformed curvature image and applying feature selection AdaBoost, with the same parameters used here, does even cause small drop in the performance. Also, there is a huge computational penalty of extracting high-level features on many reference domains and difficulty of higher-dimensionality, which makes their use prohibitive.

In fact, a well-known major benefit of feature extraction over a local support is the gain of some shift-invariance as well as illumination invariance. One must keep in mind also the fact that there is a trade-off between the size of the local support and the localization precision. However, since non-rigid registration compensates for the local deformations, wider local context cannot theoretically offer any benefit since shift-invariance has been guaranteed by registration. This can be the main reason of why high-level features do not provide any benefit in the deformable recognition scheme. In fact in a previous study investigating the relationship between features and registration in the face model-driven schemes (Chew et al., 2012), it has been shown that Gabors and HoG are not beneficial compared to pixel-based representation in the presence of non-rigid registration.

We want to reemphasize that small local context is very important, both for the deformation estimation and feature extraction parts of our algorithm. As we have clarified in Section 4.1, the mean curvature models the local context as a 2nd-order differential and it is superior to bare depth values and other types of local contexts. Furthermore, as reported in Keysers et al. (2007) where shape-model free deformable recognition of images has been investigated, local gradient context (via 3×3 Sobel filter) leads to excellent results. Therefore, in accordance with the literature, we do not suggest extraction of large local contexts as an additional analysis step following the non-rigid registration.

7.3. Spontaneous expressions and uncontrolled environments

An important current issue is the recognition of facial expressions in naturalistic environments, i.e., where the expressions are spontaneous and the environment is uncontrolled. As discussed in Corneanu et al. (2016), naturalistic environments can be characterized by varying illumination conditions, larger head poses, and low to moderate intensity of spontaneous facial expressions. Active range acquisition systems are considered quite resistant to varying illumination conditions, and 3D is beneficial for the alignment of large head poses, especially for the out-of-plane rotations.

There are recent efforts on collecting spontaneous 3D facial expression databases (Zhang et al., 2014; 2016) to be able to better address these difficulties of the spontaneous data with 3D data. Detection of subtle spontaneous expressions are especially difficult and use of the temporal representations help capturing the subtle expression differences. Therefore the recent spontaneous databases are composed of 3D video data to study the temporal aspects as well. A direct approach to capture the temporal information is to construct spatio-temporal features, as in Zhao and Pietikainen (2007) for 2D videos or as in Reale et al. (2013) for 3D videos. We refer to recent surveys for other spatio-temporal methods (Corneanu et al., 2016; Sariyanidi et al., 2015). On the other hand, more recent methods (Chu et al., 2017; Zeng et al., 2016) still use still-image feature extraction methods (e.g., local patch features like SIFT around landmarks) to cope with the hard samples of the spontaneous data, by designing person adaption mechanisms exploiting temporal information in the later stages instead of in the feature extraction stage.

While we do not propose any method to better handle the spontaneous data, in theory, our method can also be trained and evaluated on spontaneous databases as well. Although we expect to observe lower recognition rates with spontaneous data, the performance of other feature extractors would also be lower and the ranking of the tested feature extractors would not change. As shown in recent studies, such as in Eleftheriadis et al. (2015), the relative performances of different methods are quite consistent with both posed and spontaneous datasets. Therefore, we think that the databases employed in our work are sufficient to validate the potential of our novel feature extraction approach. In fact, based on its transformation invariance gained by dense registration (against transformations due to residual pose alignment errors as well as physiognomy) and also due to smaller degree of degradation observed on the difficult semi-spontaneous (SBIA) database with respect to baseline, we believe that our non-rigid registration-based feature extraction approach could be advantageous in handling the spatial variations which occur with spontaneous data. Furthermore, to better cope with the difficulties of the spontaneous data, it is possible to employ our feature extraction technique in the recent schemes like Zeng et al. (2016) and Chu et al. (2017) which depend on still-image features; or, one can envision possible extensions to use temporal dependencies in the registration process or incorporating temporal context in the recognition. Extensions of our work with temporal context and its validations on the spontaneous 3D video databases is a future research effort.

8. Conclusions

We have developed a feature extraction approach for face model-free 3D expression recognition based on a novel and effective use of non-rigid registration. This is in essence free from feature models, i.e., it is a purely data-driven recognition. Our feature extraction provides a dynamic feature extraction mechanism by forcing the facial information in a reference face gallery to adapt

to input test faces via shape model-free dense registration, which is fundamentally different from static feature extraction that obtains features always from pre-determined fixed coordinates in a reference coordinate system. This makes our simple feature representation very effective, and eliminates more complex feature representations required as in the case of static feature extraction, where the complexity arises from the necessity to model the local context, and more recently, from the deep feature hierarchies by end-to-end learning on large-scale datasets.

The proposed method is superior in two aspects: First, it enables extraction of highly informative features which are discriminative and increase inter-class distances via the merit of multiple registration references; at the same time, it attenuates the confounding variability of faces mainly due to physiognomy and residual pose misalignments. Second, right from start, our method precludes loss of information due to assumed face and/or feature models. Concomitantly, there is no need for burdensome manual face model preparation and complex model-fitting stage.

The existing shape model-free registration methods (Sotiras et al., 2013) are computationally demanding, especially because of complex regularization schemes, which makes them impractical for *expression* recognition applications in real time. On the other hand, strong *permanent face structures* make registration of 3D surfaces significantly less ill-posed, hence allow the use of simpler regularization methods. We show experimentally that our registration using local affine interpolator-based regularization is accurate when there is full correspondence and hence yields superior recognition with very fast computations (about 0.025 s per face as given in Section 5.2). However, the real challenge is the presence of *expressions*, which causes missing correspondences. Rather than resorting to a complex registration, we tackle this problem by registering test faces on multiple-references, and thus we avoid the occurrence of critical missing correspondences to a large extent. Although local failures can still occur at the fewer missing correspondences, still superior recognition performance is achieved, since only the most informative pixels from a small set of optimally selected reference domains (see Fig. 7) are picked as features.

Our comparative experiments using common high-level feature extraction techniques, like Gabor filter-banks and LBP descriptors, show the superiority of non-rigid registration-based curvature pixel features on a set of different *expression* recognition problems: recognition of basic expressions, of facial action units, and of emotional valences, using three databases with differing 3D imaging qualities. The superior advantages of our method derive from local shift-invariance on the features and the use of an optimal reference set. Local shift-invariance is enabled by non-rigid registration, which provides invariance to physiognomy (subject-invariance) and residual pose misalignments. The optimally selected face reference set improves discrimination and mitigates missing correspondence problem. A proof of the robustness of our non-rigid registration-based feature extraction is that the performance suffers very little whenever a weaker classifier, e.g., a linear classifier is used, or when the number of features is decreased. The recognition performance of other methods or with other feature types in the literature suffers much more under these conditions. Another proof of this robustness in the case of low-fidelity noisy 3D data as in the SBIA dataset.

Our approach is currently not applicable for 2D texture feature extraction since 2D texture is not directly associated with face deformations and it is affected by high degree of variations due to other factors. However, since non-rigid registration is an active research topic in face model-driven recognition, and our findings suggest its importance in model-free recognition as well, developing a suitable method for the texture data based on the proposed idea, could be the subject of a future work. We think that, whether on 3D or 2D data, non-rigid registration should also be an important aspect for future model-free recognition.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cviu.2017.07.005.

Appendix

Mutual Information for Normal Bayes. For Normal Bayes classifier (NBC), the class conditional densities are in the form of Normal densities $\{p(x|y=c) = N(x|\mu_c, \sigma_c^2\}_{c=0}^{C-1}$ where y is the class variable and C is the number of classes. If we approximate the distribution of the whole population as the Normal density, $N(x|\mu,\ \sigma^2)$, then the mutual information between x and y is evaluated as

$$I(X;Y) = H(X) - H(X|Y) \tag{13}$$

$$= \int p(x)\ln p(x)dx \tag{14}$$

$$-\sum_c p(y=c)\int p(x|y=c)\ln p(x|y=c)dx$$
$$= \frac{1}{2}\ln 2\pi e\sigma^2 - \sum_c p_c \frac{1}{2}\ln 2\pi e\sigma_c^2 \tag{15}$$

$$= \ln\sigma - \sum_c p_c \ln\sigma_c \tag{16}$$

since entropy of normal distribution is $H(X) = \frac{1}{2}\ln 2\pi e\sigma^2$ and constant terms cancel out. Here, $p_c = p(y=c)$ is the prior for class c.

We can express σ, i.e. the variance of x, in terms of means and variances of class conditional densities as follows. The moments of the whole population are

$$\mu = E[X] \tag{17}$$

$$\sigma^2 = E[(X - E[X])^2] = E[X^2] - E[X]^2. \tag{18}$$

Thus, $E[X^2] = \sigma^2 + \mu^2$. Then the mean and variance can be expressed as

$$\mu = E[X] = E[E[X|Y]] \tag{19}$$

$$= \sum_c p(y=c)E[X|Y=c] = \sum_c p_c\mu_c \tag{20}$$

$$\sigma^2 = E[X^2] - E[X]^2 = E[E[X^2|Y]] - E[X]^2 \tag{21}$$

$$= \sum_c p(y=c)E[X^2|Y=c] - E[X^2]$$
$$= \sum_c p_c(\sigma_c^2 + \mu_c^2) - \left(\sum_c p_c\mu_c\right)^2$$
$$= \sum_c p_c\sigma_c^2 + \sum_{(c,c')\in[S_C]^2} 2p_cp_{c'}\left(\mu_c^2 + \mu_{c'}^2 - 2\mu_c\mu_{c'}\right)$$
$$= \sum_c p_c\sigma_c^2 + \sum_{(c,c')\in[S_C]^2} p_cp_{c'}(\mu_c - \mu_{c'})^2 \tag{22}$$

($[S_C]^2$ is 2-combination set of $S_C = \{0, 1, \cdots, C-1\}$).

Let's use symbol V for the summation term over variances and M for the summation term over square differences of means, i.e., $\sigma^2 = V + M$. Then

$$I(X;Y) = \frac{1}{2}\ln(V+M) - \sum_c p_c \ln \sigma_c \tag{23}$$

$$= \frac{1}{2}\ln\left(1+\frac{M}{V}\right) + \frac{1}{2}\ln V - \ln\prod_c \sigma_c^{p_c} \tag{24}$$

$$= \frac{1}{2}\ln\left(1+\frac{M}{V}\right) - \ln\frac{\prod_c \sigma_c^{p_c}}{\sqrt{\sum_c p_c\sigma_c^2}}. \tag{25}$$

The second term depends only on the variances and priors, and changes insignificantly compared to the first term since it is the ratio of weighted geometric mean to weighted arithmetic mean. In particular, if the variances are all set equal, which means that the classifier is a linear classifier, the second term becomes zero. Therefore, the mutual information for NBC can be approximated by the ratio

$$I(X;Y) \approx \frac{M}{V} = \frac{\sum_{(c,c')\in[S_C]^2} p_c p_{c'}(\mu_c - \mu_{c'})^2}{\sum_c p_c\sigma_c^2} \tag{26}$$

This ratio shows that the mutual information increases when overall separation between the class means increases or when within class variances decrease. The contributions from classes are weighted by their prior probabilities. For binary classification problem, the ratio in Eq. (26) reduces to

$$I(X;Y) \approx \frac{M}{V} = \frac{p_1 p_0(\mu_1-\mu_0)^2}{p_1\sigma_1^2 + p_0\sigma_0^2} \tag{27}$$

For instance if $p_0 = p_1 = 0.5$, then

$$\frac{M}{V} = \frac{\frac{1}{4}(\mu_1-\mu_0)^2}{\frac{1}{2}\left(\sigma_1^2+\sigma_0^2\right)} = \frac{1}{4}\left(\frac{\mu_1-\mu_0}{\sqrt{\frac{1}{2}\left(\sigma_1^2+\sigma_0^2\right)}}\right)^2 \tag{28}$$

which is commonly used as discrimination or sensitivity index in signal detection theory (Green and Swets, 1966).

References

Bosphorus 3d face database URL http://bosphorus.ee.boun.edu.tr/default.aspx.

Kahou et al., S.E., 2013. Combining modality specific deep neural networks for emotion recognition in video. ICMI.

Berretti, S., Del Bimbo, A., Pala, P., Amor, B.B., Daoudi, M., 2010. A set of selected SIFT features for 3D facial expression recognition. ICPR.

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3d faces. In: ACM SIGGRAPH, pp. 187–194.

Chew, S.W., Lucey, P., Lucey, S., Saragih, J.M., Cohn, J.F., Matthews, I., Sridharan, S., 2012. In the pursuit of effective affective computing: the relationship between features and registration.. IEEE Trans. Syst. Man Cybern. Part B 42 (4), 1006–1016.

Chu, W.S., l. Torre, F.D., Cohn, J.F., 2017. Selective transfer machine for personalized facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. 39 (3), 529–545.

Corneanu, C.A., Oliu, M., Cohn, J.F., Escalera, S., 2016. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. 38 (8), 1548–1568.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. Signal Process. Mag. IEEE January (1), 32–80.

Creusot, C., Pears, N., Austin, J., 2013. A machine-learning approach to keypoint detection and landmarking on 3d meshes. Int. J. Comput. Vis. 102 (1–3), 146–179.

Dahmane, M., Meunier, J., 2011. Emotion recognition using dynamic grid-based hog features. In: IEEE FG.

Dhall, A., Asthana, A., Goecke, R., Gedeon, T., 2011. Emotion recognition using phog and lpq features. In: IEEE FG Workshops.

Ding, H., Zhou, S. K., Chellappa, R., 2016. FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. arXiv:1609.06591.

Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. 17 (2), 124–129.

Ekman, P., Friesen, W.V., Hager, J.C., 2002. Facial Action Coding System. A Human Face, Salt Lake City, UT.

Eleftheriadis, S., Rudovic, O., Pantic, M., 2015. Multi-conditional latent variable model for joint facial action unit detection. In: IEEE ICCV.

Fang, T., Zhao, X., Ocegueda, O., Shah, S.K., Kakadiaris, I.A., 2012. 3D/4D facial expression analysis: an advanced annotated face model approach. Image Vis. Comput. 30 (10), 738–749.

Gehrig, T., Ekenel, H., 2011. Facial action unit detection using kernel partial least squares. In: IEEE ICCV Workshops.

Gray, A., 1997. Modern Differential Geom. of Curves and Surfaces with Mathem., 2nd CRC Press, Inc., Boca Raton, FL, USA.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.

Jaiswal, S., Valstar, M., 2016. Deep learning the dynamic appearance and shape of facial action units. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8.

Jung, H., Lee, S., Yim, J., Park, S., Kim, J., 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE ICCV.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1867–1874.

Keysers, D., Deselaers, T., Gollan, C., Ney, H., 2007. Deformation models for image recognition. IEEE Trans. Pattern Anal. Mach. Intell. 29 (8), 1422–1435.

Khorrami, P., Paine, T.L., Huang, T.S., 2015. Do deep neural networks learn facial action units when doing expression recognition? In: IEEE ICCV Workshops.

Kim, B.-K., Lee, H., Roh, J., Lee, S.-Y., 2015. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In: ACM ICMI.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Levi, G., 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ACM ICMI.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M., 2011. The computer expression recognition toolbox (cert). In: IEEE FG, pp. 298–305.

Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A., Berretti, S., 2011. Shape analysis of local facial patches for 3d facial expression recognition. Pattern Recognit. 44 (8), 1581–1589.

Mpiperis, I., Malassiotis, S., Strintzis, M., 2008. Bilinear elastically deformable models with application to 3d face and facial expression recognition. IEEE FG. Amsterdam.

Ng, H.-w., Nguyen, V.D., Vonikakis, V., Winkler, S., 2015. Deep learning for emotion recognition on small datasets using transfer learning. In: ACM ICMI.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: British Machine Vision Conference.

Ramanathan, S., Kassim, A.A., Venkatesh, Y.V., Wah, W.S., 2006. Human facial expression recognition using a 3d morphable model. IEEE ICIP.

Reale, M., Zhang, X., Yin, L., 2013. Nebula feature: a space-time feature for posed and spontaneous 4d facial behavior analysis. In: IEEE FG.

Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 FPS via regressing local binary features. In: Proc. IEEE CVPR, 1 (1), pp. 1685–1692.

Rusu, R., Cousins, S., 2011. 3d is here: point cloud library (pcl). IEEE ICRA.

Sandbach, G., Zafeiriou, S., Pantic, M., 2012. Binary pattern analysis for 3d facial action unit detection. BMVC. Guildford, UK.

Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L., 2012. Static and dynamic 3d facial expression recognition: a comprehensive survey. Image Vis. Comput. 30 (10), 683–697.

Sandbach, G., Zafeiriou, S., Pantic, M., 2012. Local normal binary patterns for 3d facial action unit detection. In: IEEE ICIP. Orlando, FL, USA, pp. 1813–1816.

Sariyanidi, E., Gunes, H., Cavallaro, A., 2015. Automatic analysis of facial affect: a survey of registration, representation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 99, 1.

Savran, A., 2011. Non-Rigid Registration-Based Data-Driven 3D Facial Action Unit Detection. Boğaziçi University Ph.D. thesis. URL http://theses.eurasip.org/theses/354/non-rigid-registration-based-data-driven-3d/.

Savran, A., Gur, R., Verma, R., 2013. Automatic detection of emotion valence on faces using consumer depth cameras. IEEE ICCV Workshops. Sydney, Australia.

Savran, A., Sankur, B., 2008. Non-rigid registration of 3d surfaces by deformable 2d triangular meshes. IEEE CVPR Workshops. Anchorage, Alaska, USA.

Savran, A., Sankur, B., 2009. Automatic detection of facial actions from 3d data. IEEE ICCV Workshops. Kyoto, Japan.

Savran, A., Sankur, B., Bilge, M.T., 2012. Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. Pattern Recognit. 45 (2), 767–782.

Savran, A., Sankur, B., Bilge, M.T., 2012. Regression-based intensity estimation of facial action units. Image Vis. Comput. 30 (10), 774–784.

Shan, C., Gong, S., McOwan, P.W., 2009. Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. 27 (6), 803–816.

Sikka, K., Wu, T., Susskind, J., Bartlett, M., 2012. Exploring bag of words architectures in the facial expression domain. In: ECCV Workshops, 7584, pp. 250–259.

Sinha, A., Bai, J., Ramani, K., 2016. Deep Learning 3D Shape Surfaces Using Geometry Images. Springer International Publishing, Cham, pp. 223–240.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE Trans. Med. Imaging 32 (7), 1153–1190.

Soyel, H., Demirel, H., 2007. Facial expression recognition using 3d facial feature distances. ICIAR. Montreal, Canada.

Sun, Y., Reale, M., Yin, L., 2008. Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. IEEE FG. Amsterdam, Netherlands.
Sun, Y., Wang, X., Tang, X., 2014. Deep learning face representation from predicting 10,000 classes. In: IEEE CVPR.
Szegedy, C., Toshev, A., Erhan, D., 2013. Deep neural networks for object detection. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 2553–2561.
Tang, H., Huang, T., 2008. 3d facial expression recognition based on automatically selected features. IEEE CVPR Workshops. Anchorage, Alaska, USA.
Tsalakanidou, F., Malassiotis, S., 2010. Real-time 2d+3d facial action and expression recognition. Pattern Recognit. 43 (5), 1763–1775.
Vretos, N., Nikolaidis, N., Patras, I., 2011. 3d facial expression recognition using Zernike moments on depth images. ICIP.
Wang, J., Yin, L., Wei, X., Sun, Y., 2006. 3d facial expression recognition based on primitive surface feature distribution. IEEE CVPR. Washington, DC, USA.
Wu, T., Butko, N.J., Ruvolo, P., Whitehill, J., Bartlett, M.S., Movellan, J.R., 2012. Multilayer architectures for facial action unit recognition. IEEE Trans. Syst. Man Cybern. Part B 42 (4), 1027–1038.
Yang, S., Bhanu, B., 2012. Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans. Syst. Man Cybern. Part B 42 (4), 980–992.
Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Learning face representation from scratch. CoRR abs/1411.7923.
Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J., 2006. A 3d facial expression database for facial behavior research. IEEE FG. Southampton, UK.
Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. Springer International Publishing, Cham, pp. 818–833.
Zeng, J., Chu, W.S., la Torre, F.D., Cohn, J.F., Xiong, Z., 2016. Confidence preserving machine for facial action unit detection. IEEE Trans. Image Process. 25 (10), 4753–4767.
Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M., 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image Vis. Comput. 32 (10), 692–706. Best of Automatic Face and Gesture Recognition 2013
Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L., 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3438–3446.
Zhao, G., Pietikainen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 915–928.
Zhao, K., Chu, W., la Torre, F.D., Cohn, J.F., Zhang, H., 2015. Joint patch and multi-label learning for facial action unit detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp. 2207–2216.
Zhao, K., Chu, W.S., Zhang, H., 2016. Deep region and multi-label learning for facial action unit detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3391–3399.
Zhao, X., Dellandra, E., Zou, J., Chen, L., 2013. A unified probabilistic framework for automatic 3d facial expression analysis based on a Bayesian belief inference and statistical feature models. Image Vis. Comput. 31 (3), 231–245.