Krešimir Bešenić[1], Ivan Gogić[2], Igor S. Pandžić[2] and Krešimir Matković[3]

Automatic Image-based Face Analysis Systems Overview

[1]Visage Technologies, Diskettgatan 11A, SE-583 35 Linköping, Sweden
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
[3]VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Donau-City-Straße 11, 1220 Wien, Austria

Abstract

Face analysis systems have recently gained popularity due to the large number of potential applications across a wide range of industries. Various types of information can be extracted from an image of a face including: face location and size, location of characteristic facial landmark points, 3D head pose, facial expression and emotion, gaze direction and biometric information (i.e. age, gender and race). Most of these problems are solved using machine learning techniques based on large sets of training samples. Furthermore, information from these different tasks is often complementary and can be used to enhance the accuracy of the algorithms. A systematic overview of current approaches to face analysis tasks is presented as an introduction to this growing research field.

1. Introduction

Applications of Face Analysis (FA) technologies span numerous and diverse industrial and commercial fields.

Currently the most common applications are found in **marketing and entertainment**, based on the novelty and fun effect of FA, usually combined with 3D graphics, such as the popular face masks in Snapchat and similar apps. Numerous major brands have used FA effects in their online marketing campaigns. Furthermore, products such as make-up, glasses or even hats and earrings, use **virtual try-on** applications for promotion and testing. Many such applications allow direct purchasing. In physical **retail** spaces, experiments are starting to analyze customer behavior and shopping patterns using cameras placed in shops or shop windows. In **marketing research**, analysis of subjects' gaze patterns and emotional reactions has traditionally been performed in on-site studies using specialized gaze tracking hardware and requiring large number of subjects. The new generation of marketing research technology uses FA software to perform similar research with subjects participating from home (being paid as micro-workers), dramatically reducing cost and increasing speed and scale of possible research. **Automotive** industry has been deploying various forms of fatigue detection in heavy commercial vehicles and, more recently, in cars. However, the use of FA for driver monitoring is still in fairly early stages and we expect to see much more widespread deployment in years to come. Furthermore, there is interest in other uses of FA such as controlling the information system or automatic personal adjustments in high-end cars. By monitoring operators of various types of machinery (e.g. forklifts), FA can help increase **industrial safety**. **Assistive technologies** help people with disabilities perform various tasks by using limited movement such as gaze or head motion. **Biometrics** based on face recognition is increasingly deployed for access control (e.g. to financial services). To avoid trivial fraud by submitting a picture instead of the live face, such applications deploy liveness detection techniques based on FA. Ubiquitous computing power and variety of available sensors are already changing the way we treat **health**, allowing simplified and more widespread monitoring and diagnostics through inexpensive devices and apps. FA technologies play a role in this trend, with experimental or prototype applications for remote fever detection, posture monitoring, concussion diagnostics and others. Further applications of FA include **robotics**, where it allows robots to interact with humans, and advanced **audio** systems that use 3D head position to deliver perfect sound to the listener.

2. General Face Analysis Framework

A typical face analysis framework can be viewed as a pipeline consisting of several steps. As in many other image analysis frameworks, the first step is object (face) detection. The face detection step is usually followed by preprocessing, face alignment, feature extraction, and attribute prediction steps, sequentially (Figure 1).

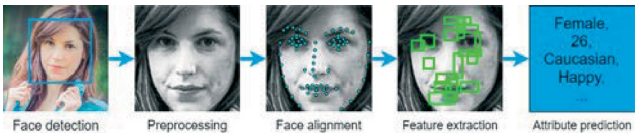

Fig. 1. General face analysis framework

2.1. Face Detection

In the proposed framework pipeline, initial face detection step is semi-decoupled from other steps as it results in basic face location and scale information, usually provided via a facial bounding box. Even though the bounding box information is most basic, different versions of face detectors can be trained with differently defined bounding boxes and can results in different detection qualities, thus introducing bias and propagating the error to the rest of the pipeline. To some extent, this can be alleviated in a preprocessing step and by introducing perturbation augmentations to the training set. Most widely used face detection systems are based on the work of Viola and Jones [1] and, more recently, deformable parts models [2] and single shot detection systems [3]. More details on face detection methods can be found in [4]

2.2. Preprocessing

Depending on the face analysis method, the preprocessing step can be as trivial as image cropping based on the facial bounding box. Typical minimal preprocessing techniques also include resizing and color conversions. To deal with low contrast and lighting problems, additional preprocessing techniques such as histogram equalization, Difference of Gaussians filtering, and edge enhancement filtering (e.g. Sobel filtering) can be incorporated.

2.3. Face Alignment

To compensate for face detector inaccuracies and to deal with misaligned faces captured in unconstrained conditions, various alignment techniques have been proposed. Most basic method rests on face detection confidence. The input image is rotated by a small angle multiple times and the version with the highest detection confidence is used. Although computationally expensive, this simple method does not introduce any new components (existing face detection system is reused) and can result in a satisfying performance. More complex methods rely on facial landmark point detection methods. Given a set of detected landmark points, in-plane rotation and scaling can be performed based on eye points locations, Procrustes Analysis transform or 3D face model fitting.

2.4. Feature extraction

It is well known that any classifier is only as good as the data it works with. This applies to all types of face analysis systems, therefore in many cases making the feature extraction step the most important one. Geometric feature extraction, which is based on fiducial distance measurements, heavily relies on precise facial landmark points detection. Geometric features can be reliable in 3D use-cases, yet in the case of 2D images, their practicality is usually restricted to constrained frontal neutral faces. Appearance-based features consist of pixel values or their transformations, thus making them more suitable for 2D image use. While raw pixel values can be used directly as an input for classification and regression systems, more elaborate approaches such as Local Binary Patterns (LBP), Biologically Inspired Features (BIF), Haar-like features, Histogram of Oriented Gradients (HOG) features, speeded up robust features (SURF), and Gabor filters are commonly used.

2.5. Attribute prediction

In this work, attribute prediction refers to classification and regression tasks related to expression, age, gender, and race prediction. Binary classification is commonly used for gender classification and simple race classifiers (e.g. Asian/Non-Asian, White/Black). Multi-class classification is used for face expression and age group classifiers, and in some cases for exact age prediction (e.g. 100 classes, one for each year). Regression is a natural (but not necessarily optimal) choice for exact age estimation. Han et al. combined classification and regression in their proposed hierarchical estimator consisting of a between-group classification and a within-group regression in [5].

Some methods, most notably neural networks, perform multiple steps in a joint manner. Deep Convolutional Neural Networks (DCNN) combine feature extraction and attribute prediction steps together to learn optimal feature extractors and model high-level abstractions from the data. Due to their flexibility, DCNNs can be used for classification, regression or for more elaborate combinations of those two approaches.

3. Face Alignment

Face alignment is the process of determining the location of characteristic facial features or landmarks (points that delineate eyes, nose, mouth, eyebrows, chin and face contour) given a face image. The configuration of facial landmarks is also usually referred to as face shape which is represented as a vector of 2D landmark coordinates. Various machine learning algorithms are employed in order to estimate the face shape. If we denote it with $S = (x_1, y_1,..., x_L, y_L)$ where L represents the number of landmarks, the goal of face alignment, given a face image, is to find a shape S closest to the ground truth shape S^*. More formally, the goal is to minimize:

$$||S - S^*|| \quad (1)$$

where $||\cdot||$ is a suitable vector norm. The alignment error in (1) is used as a performance measure that drives the training process.

Regression methods estimate the face shape directly from image features and have recently demonstrated superior accuracy, speed and robustness when compared to earlier, traditional methods that involve Active Appearance Models, Active Shape Models and local part classification using search algorithms. Such constructed models demonstrate poor ability to express all combinations of face variations due to expressions, illumination and head pose [6].

Regression methods can be roughly divided into four categories: constrained regression, cascaded regression, deep learning, and head pose and occlusion methods. Constrained regression methods estimate landmark positions individually, then additionally ensure a probable face configuration. However, in cascaded regression framework, an implicit face shape constraint is incorporated into the training process. This framework is currently the standard approach to face alignment. Recently, with advances in computing power and optimization techniques, Convolutional Neural Networks (CNN) have been applied to face alignment as part of the deep learning category. With the growing success of cascaded regression and deep learning methods, face alignment in more challenging conditions has become the focus area for researchers as part of the head pose and occlusion category.

Cascaded regression has established itself as the leading approach for face alignment due to its speed, robustness and accuracy. In this framework, a number of regressors $(R_1 ,...,R_t ,...,R_T)$ are successively applied starting from the initial shape estimate S_0. Given an image I, each regressor learns and estimates a shape increment δS and updates the face shape:

$$\delta S = R_t (I, S_{t-1}) \quad (2)$$

$$S_t = S_{t-1} + \delta S \quad (3)$$

where the tth regressor R_t updates the previous shape S_{t-1} to the new shape S_t (Cao et al. 2014). It is important to note that the tth regressor depends on the previous shape estimate S_{t-1}. The dependency is usually through shape-indexed features which is a concept first introduced in [7]. These features are stored relative to the object pose and are thus consistent across large pose variations.

4. Facial Expression Recognition

In order to automatically recognize emotions and their related expressions, an investigation on how to define those terms needed to be done first. In [8], Ekman and Friesen discovered six basic or prototypic emotions (anger, disgust, fear, happiness, sadness, and surprise) whose facial expressions are culturally and racially invariant and are, therefore, great candidates for automatic systems which need clear categories. However, one important drawback of this model became evident. It is too crude to accurately model the complexity of emotions people experience in everyday lives. As a response, Facial Action Coding System (FACS) was developed in order to define atomic facial muscle movements named Action Units (AU) spanning the whole spectrum of human facial expressions. Its aim is objectivity in the signal measurement which is separated from the final expression classification often influenced by the context. Consequentially, a group of researchers tried to develop algorithms that recognize these simpler, intermediate categories and synthesize the final expression afterward. However, FACS annotation is a very tedious process which requires expert knowledge few people poses. Therefore, few data sets with full FACS annotations are available to the community. The six basic expressions classification approach is currently the most widely used categorization in computer vision.

The features used for Facial Expression Recognition (FER) can roughly be divided into appearance and geometric-based. The appearance features are extracted from facial image intensities to represent a discriminative textural pattern while the geometric ones need accurate landmark positions from which different relations can be constructed. The geometric features are, however, very sensitive to the individual face shape configuration and are therefore less consistent in person independent scenarios.

Well known and widely successful hand-crafted features such as variations of Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), Gabor filters and Local Phase Quantization (LPQ) descriptors have also been considered for FER. While most approaches considered a regular grid of patches or the whole face region for feature extraction, there have been advances in determining common and specific salient facial re-

gions for each expression. In [9], Happy and Routray demonstrated the importance of facial landmark detection in order to find the salient patches from which they extract features.

On the other hand, a number of researchers tried to fuse different texture encoding features in order to extract complementary information that would benefit the FER. For instance, Zhang et al. used multiple kernel learning to combine two feature representations: HOG and LBP [10].

While all of the previously mentioned methods use hand-crafted and heuristically determined features, experiments with deep learning using CNN on the FER problem were recently conducted as well. However, deep learning methods have serious over-fitting problems with small datasets that are typical for FER. Lopes et al. tried different preprocessing techniques (image normalizations, synthetic samples etc.) in order to cope with the mentioned problem and were able to achieve state-of-the-art results on the CK+ benchmark dataset [11]. Even though real-time performance is claimed, a high-end GPU is needed in order to achieve it.

An additional direction of research is to integrate temporal dimension into both appearance and geometric features when working with image sequences.

5. Biometric Attributes Estimation

Biometrics refers to the problem of subject identification based on a certain unique physical characteristic (i.e. fingerprint, iris or face). On the other hand, soft biometric attributes are traits such gender, height, and eye color that provide some useful information about the subject, but are not distinctive enough to perform identification [12]. The intrusiveness of biometric systems based on fingerprint or iris recognition reduces their applicability compared to systems based on facial image analysis that do not require physical contact, subject's cooperation nor subject's attention. Three most prominent and widely researched soft biometric attributes that can be estimated from facial images are gender, age and race.

5.1. Gender classification

Gender classification is a fundamental soft biometric attribute estimation task. Due to the significance of gender attribute, availability of public face datasets with gender labels, and simplicity of the task itself (binary classification), it was a recurrent topic in early work on facial analysis.

In 1991., human performance was matched by a simple multi-layer perceptron system named SexNet [13]. By directly using pixel values as features, they achieved accuracy of 91.9% on a manually collected dataset containing 90 images.

Discriminative properties of 7 different facial regions were evaluated on a dataset containing 800 frontal faces in [14]. Periocular region was shown to be the most informative and their multi-region method based on upper face region, left eye, and nose yielded a 5% lower classification error compared to the holistic face approach.

A step further was taken in [15]. Local discriminative DNNs were applied to the most informative facial regions determined by Sobel filtering, blurring and binarization. Experiments were performed on the aligned version of LFW dataset with 13,233 images and a subset of the even more difficult Groups dataset containing 14,760 images. Evaluation on large unconstrained datasets demonstrated in-the-wild effectiveness and cross-database experiments verified the generalization capability of the proposed approach.

Despite using a simpler holistic-face approach, previously mentioned methods were outperformed by a straight-forward CNN approach trained on 500k images [16]the problem of gender recognition from face images remains difficult when dealing with unconstrained images in a cross-dataset protocol. In this work, we propose a convolutional neural network ensemble model to improve the state-of-the-art accuracy of gender recognition from face images on one of the most challenging face image datasets today, LFW (Labeled Faces in the Wild, demonstrating the power of CNNs.

5.2. Age estimation

Age estimation is one of the most challenging and broadly researched topics in the face analysis field. Early work was primarily based on geometric features, ageing pattern subspaces or manifold learning. A drawback of the mentioned approaches is that they require a well-aligned frontal faces. This section focuses on appearance-based methods that are more suited for unconstrained faces.

Age estimation can be viewed as a multi-class classification or a regression problem. Recently, the label distribution method is frequently used as it combines best of the two approaches. Age estimation typically refers to the real (chronological, biological) age estimation. Apparent age estimation is a more recent endeavor, referring to age estimation as perceived by other humans. Apparent age estimators are trained on datasets where there is no ground truth real age but instead, a group of people was guessing the subject's age.

A seminal real age estimation method based on Biologically Inspired Features (BIF) was proposed in [17]. Their variation of BIF used Gabor filters to model receptive fields and MAX and STD operations as sources of nonlinearity. PCA was used for dimensionality reduction, followed by a linear SVM in case of age classifi-

cation or a support vector regressor (SVR) in case of age regression experiments. Their approach outperformed all previous work on the FG-NET benchmark, and was a basis for a number of BIF-related age estimation methods.

Comparison of hand-crafted and learned features under the same experimental settings was performed in [18]. Their exhaustive experiments showed that a simple CNN with only 2 feature-extracting convolutional layers can outperform different combinations of hand-crafted features (i.e. HOG, LBP and SURF).

Power of CNNs was further demonstrated in [19]. Their age estimation approach was based on an ensemble of 20 VGG-16 models pretrained for the task of image classification. The networks were trained for classification with 101 output neurons, each corresponding to an age from interval 0-100. The final prediction was the softmax-normalized output of those neurons, averaged over the 20 networks. An impressive error drop was achieved by additional pretraining on their large and noisy IMDB-WIKI dataset, improving the state-of-the-art by a large margin. By performing additional fine-tuning on the apparent age LAP dataset, they also achieved top score on the first edition of the ChaLearn apparent age challenge.

5.3. Race classification

In the face analysis research field, terms ethnicity and race are often used interchangeably. However, they are related to sociological and biological factors respectively. Generally, ethnicity is viewed as a cultural concept, while race refers to the person's physical appearance or characteristics and is a better suited term for classification based on facial images. Categorization to 7 commonly accepted racial groups covers more than 95\% of the world population. However, due to the scarcity of public datasets with racial annotations and good sample distribution, most of the race classification research is done on simple binary (e.g. Asian/non-Asian, White/Black) or ternary (e.g. Caucasian/African American/Asian) classification.

Following the success of deep neural networks in many other face analysis fields, Wang et al. [20] showed superior performance of their DCNN method on both binary and ternary race classification tasks. Their approach was based on CIFAR-10 CNN architecture with a n-way softmax layer. A cross-entropy loss was used during the training and the networks were trained for 3 different scenarios: (i) classification of White and Black subjects, (ii) classification of Chinese and Non-Chinese subjects, and (iii) classification of Han, Uyghur, and Non-Chinese subjects. To deal with the lack of public large-scale race analysis databases, they worked with different combinations of several public face analysis datasets and additional private datasets. For all 3 scenarios, they reported classification accuracies from 99.4% to 100%.

6. Conclusion

Recent progress in development of Face Analysis (FA) technologies created an opportunity for many new innovative commercial application fields. Like in many other computer vision fields, trend towards adoption of CNNs and deep learning is obvious, but in many cases inference speed and memory requirements are neglected. Additionally, the lack of dedicated large-scale datasets becomes more obvious due to the overfitting problems. In most cases, FA methods focus on estimation of a single attribute from a single image. Integrating the temporal dimension and solving multiple tasks jointly could increase the algorithms performance.

References

[1] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[2] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," *Lect. Notes Comput. Sci.*, vol. 8692 LNCS, no. PART 4, pp. 720–735, 2014.

[3] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci.*, vol. 9905 LNCS, pp. 21–37, 2016.

[4] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," *Microsoft Res.*, no. June, p. 17, 2010.

[5] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.

[7] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1078–1085, 2010.

[8] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[9] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[10] X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using lp -norm MKL multiclass-SVM," *Mach. Vis. Appl.*, vol. 26, no. 4, pp. 467–483, May 2015.

[11] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.

[12] C. B. Ng, Y. H. Tay, and B. M. Goi, "A review of facial gender recognition," *Pattern Anal. Appl.*, vol. 18, no. 4, pp. 739–755, 2015.

[13] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "Sexnet: A neural network identifies sex from human faces," *Adv. Neural Inf. Process. Syst. 3*, no. July, pp. 572–7, 1991.

[14] L. L. L. Lu, Z. X. Z. Xu, and P. S. P. Shi, "Gender Classification of Facial Images Based on Multiple Facial Regions," *WRI World Congr. Comput. Sci. Inf. Eng.*, vol. 6, pp. 48–52, 2009.

[15] J. Mansanet, A. Albiol, and R. Paredes, "Local Deep Neural Networks for gender recognition," *Pattern Recognit. Lett.*, vol. 70, pp. 80–86, 2016.

[16] G. Antipov, S. A. Berrani, and J. L. Dugelay, "Minimalistic CNN-based ensemble model for gender prediction from face images," *Pattern Recognit. Lett.*, vol. 70, pp. 59–65, 2016.

[17] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, pp. 112–119, 2009.

[18] I. Huerta, C. Fernández, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," *Pattern Recognit. Lett.*, vol. 68, pp. 239–249, 2015.

[19] R. Rothe, R. Timofte, and L. van Gool, "Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks," *Int. J. Comput. Vis.*, pp. 1–14, 2016.

[20] W. Wang, F. He, and Q. Zhao, "Biometric Recognition," vol. 9967, pp. 176–185, 2016.