Signal Processing 178 (2021) 107755

Contents lists available at ScienceDirect

Signal Processing

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

Review

Regression-based methods for face alignment: A survey

Ivan Gogić[a,*], Jörgen Ahlberg[b], Igor S. Pandžić[a]

[a] Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia
[b] Dept. of Electrical Engineering, Computer Vision Laboratory, Linköping University, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 29 October 2019
Revised 9 June 2020
Accepted 7 August 2020
Available online 14 August 2020

Keywords:
Face alignment
Facial feature localization
Facial landmarks detection
Survey
Regression

ABSTRACT

Face alignment is the process of determining a face shape given its location and size in an image. It is used as a basis for other facial analysis tasks and for human-machine interaction and augmented reality applications. It is a challenging problem due to the extremely high variability in facial appearance affected by many external (illumination, occlusion, head pose) and internal factors (race, facial expression). However, advances in deep learning combined with domain-related knowledge from previous research recently demonstrated impressive results nearly saturating the unconstrained benchmark data sets. The focus is shifting towards reducing the computational burden of the face alignment models since real-time performance is required for such a highly dynamic task. Furthermore, many applications target devices on the edge with limited computational power which puts even greater emphasis on computational efficiency. We present the latest development in regression-based approaches that have led towards nearly solving the face alignment problem in an unconstrained scenario. Various regression architectures are systematically explored and recent training techniques discussed in the context of face alignment. Finally, a benchmark comparison of the most successful methods is presented, taking into account execution time as well, to provide a comprehensive overview of this dynamic research field.

1. Introduction

Face alignment is the process of determining the face shape, i.e. the location of characteristic facial features or landmarks (points that delineate eyes, nose, mouth, eyebrows, chin, and face contour) given a face image (Fig. 1). A vast majority of face alignment methods assume that the face bounding box is known both at training and testing phases. The face bounding box is usually obtained through face detection algorithms (see [1]) or from manual annotations (“ground truth”). 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 $S = (x_1, y_1, \ldots, x_L, y_L)$ 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:

$$||S - S^*|| \tag{1}$$

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

Face alignment is a well studied area of computer vision. It is often used as a stepping stone for other important tasks in face analysis such as emotion and expression recognition [2,3], face recognition [4,5], and face tracking [6]. More specifically, facial landmark points are used to extract salient features helping facial expression recognition algorithms to focus on relevant regions of the face [7]. For face recognition, face alignment is necessary as a preprocessing step to register and align facial images in order to eliminate in-plane rotations and provide consistent facial crops for further processing [8]. Finally, detected facial landmark points are the key component to correctly estimate the parameters of a 3D Morphable Model (3DMM) [9] providing 3D head pose and facial action units as a result.

Face alignment has wide application areas in many different industries, including human-machine interaction, video conferencing, gaming, animation, and augmented reality. The applications range from fun augmented-reality gimmicks such as face masking or virtual make-up, to life-saving technology in automotive industry like driver distraction and drowsiness detection. For all these reasons, it rightfully received attention from the computer vision research community.

The purpose of this work is to evaluate and give an overview of the most successful machine learning techniques that solve this complex task. Unlike the recent, more general review from Jin and Tan [10] which covers the breadth of different approaches from the

[*] Corresponding author.
E-mail addresses: ivan.gogic@fer.hr (I. Gogić), jorgen.ahlberg@liu.se (J. Ahlberg), igor.pandzic@fer.hr (I.S. Pandžić).

https://doi.org/10.1016/j.sigpro.2020.107755
0165-1684/© 2020 Elsevier B.V. All rights reserved.

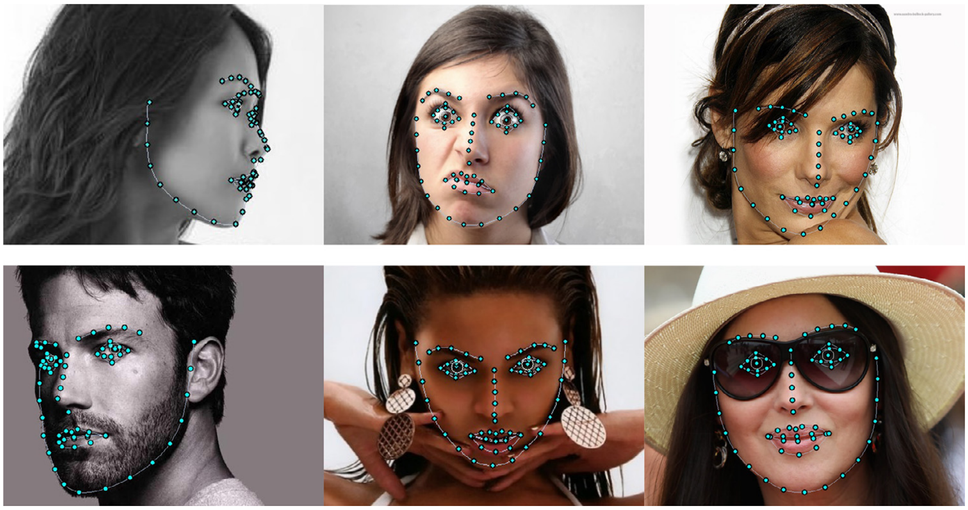


Fig. 1. Examples of face alignment in large variations of head pose, occlusion level, expression, and illumination.

last 20 years, we focus on regression methods which estimate the face shape directly from image features. These methods demonstrated superior accuracy, speed, and robustness when compared to earlier, traditional methods that involve Active Appearance Models [11,12], Active Shape Models [13], 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 [14]. The situation has changed and we are no longer far from solving the face alignment "in-the-wild" as reported in [10] due to the latest developments in regression-based approaches. To summarize, the main contributions of this work are:

- Systematic overview of regression-based/discriminative methods outlining their evolution towards human-level accuracies on challenging data sets.
- Comprehensive comparison on relevant "in-the-wild" benchmark data sets including computational efficiency.
- Insights for future research directions extracted from the comparison of the most successful methods.

An overview of regression architectures and its variants is given in Section 2, highlighting the strengths and weaknesses of each approach. In Section 3, 3D alignment methods are introduced and analyzed. Multi-task learning, a relatively new technique, is investigated in Section 4 in the context of face alignment. Another important topic for face alignment is described in Section 5, exploring how to handle partial occlusions of the face. Finally, the presented methods are compared on relevant benchmark data sets taking execution time into account in Section 6 with final conclusions in Section 7.

2. Regression architectures

Regression-based or, as they are also often called, discriminative methods, estimate landmark positions directly from facial images. It is usually formulated as a standard regression problem where the target values are difference vectors between an initial shape estimate and the ground truth shape using features extracted from images. The initial shape estimate is usually a mean shape calculated from the training set normalized to the ground truth bounding box.

Earlier methods used regression for each landmark individually based on the local appearance around the initial position and additionally enforced a global shape constraint to make the local estimations more robust. These methods are described in more detail in Section 2.1. Later on, researchers used the joint training process for all landmarks to create an implicit shape constraint making the methods more straight-forward and simple to train. The innovation came in pair with a new cascaded architecture which breaks up the problem and solves it in a coarse-to-fine manner. This cascaded architecture achieved good results and has been further developed into many variants which are systematically covered in Sections 2.2, 2.3 and 2.4. Finally, as in many other computer vision fields, researchers investigated deep learning methods and Convolutional Neural Networks (CNNs) for face alignment both in a pure end-to-end and cascaded regression framework (Section 2.5).

2.1. Constrained regression

Constrained regression methods learn to induce individual landmark positions directly from image features but also employ a corrective step which ensures global face shape constraint.

A representative algorithm from this group is Boosted Regression with Markov Networks (BoRMaN) [15]. Support Vector Regression (SVR) with a Gaussian Radial Basis Function (RBF) kernel is used as a local regressor for each landmark. The method uses Haar filter responses as features. An initial estimate of landmark locations is a mean shape placed relative to a bounding box returned by a face detector. Each prediction is then refined using Markov Random Fields (MRF) that model the global relations between landmarks. Each node in the model is a vector between two landmarks. The relation between two nodes is modeled as the difference of angles and the ratio of the lengths of these vectors (Fig. 2). This ensures the robustness to scale, rotation, and translation variations. The positions of the landmarks are iteratively updated. At each iteration the Markov Network analyzes the current predictions and generates the sampling regions for the next iteration. In the process, stable points are used to aid in the prediction of non-stable landmarks.

Kazemi and Sullivan in [16] proposed a method which uses a sliding window approach to detect the face parts (mouth, nose, and eyes) within the previously detected face region with a constructed tree structure to enforce shape constraints. After the parts are located, individual landmark linear regressors are used on the image patches to find the landmark points of the corresponding parts. A variant of Pyramid of Histograms of Orientation Gradients (PHOG) feature descriptors described in [17] are used for both detectors and regressors.

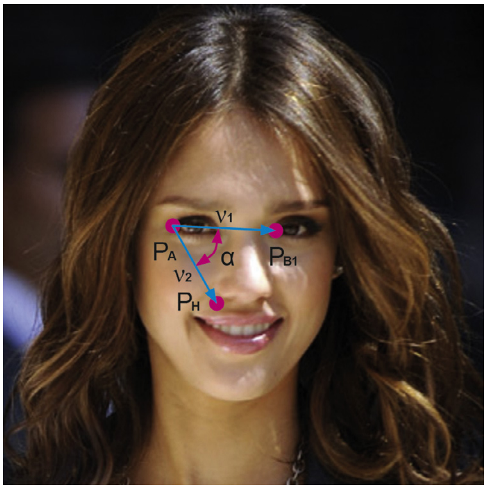


Fig. 2. The relationship between landmarks as modeled in the BoRMaN method [15].

A constraint technique similar to BoRMaN [15] is employed in [18]. The method is called Structured Output Regression Forest (SO-RF) and its spatial constraints are manually modeled by a structure graph. Each landmark has a directed graph associated with it that defines its influence to the neighboring landmarks. Each leaf node models the affiliation to a base landmark and stores an offset and a confidence. Additionally, each leaf node models the relative offsets to the neighboring landmarks with a Gaussian distribution. The combined votes from the local evidence and the spatial constraints form a map where the highest probability landmark position is found. In their later work (see [19]), Yang and Patras use the same regression forest voting scheme, but the shape constraints are replaced with sieves that act as filters for the votes. The forest leaves cast votes for the landmarks and face center simultaneously. A Hough map is formed from the votes and the mean-shift algorithm is then used to find the maximum likelihood detections for the landmarks. The first sieve used is a face center sieve that discards the votes not consistent with the global face center hypothesis. The votes are then filtered by proximity threshold sieves where the threshold is iteratively adapted based on the decision from a classifier trained on features extracted from Hough maps.

The final representative method from this group is Local Evidence Aggregated Regression (LEAR) [20]. The overall idea of the method is to use predictions from local individual regressors and shape constraints as in BoRMaN [15] to update the sampling region in the next iteration. Additionally, each iteration prediction is accumulated into a probability map from which the final prediction is made. Local Binary Patterns (LBP) [21] extracted from patches are used as feature vectors and SVR to regress the offset vector. The regressors are trained to be precise as opposed to general by limiting the variance of the training set sampling locations. The outlier predictions thus produced in inference phase are then mitigated by aggregation of all estimates from previous iterations. The regressor output is evaluated by performing another regression from the predicted location using the output distance to measure confidence.

2.1.1. Summary

The methods described in this section are the earliest attempts of robust face alignment (Table 1). It became evident that local landmark appearance, although very important, is not sufficient for accurate localization. Information from neighboring landmarks and global face shape configuration is equally important in order to solve extreme variations in facial appearance. The first attempts to utilize the face shape is through constructed constraints and corrective post-processing after individual landmark localization. It was an important step in the right direction, however, it is very difficult to manually construct such a constraint to accommodate all possible variations and still provide needed robustness. An additional weakness is the use of hand-crafted features for landmark localization that suffer from similar problems. As in other computer vision fields, a shift towards data driven modeling occurred in face alignment as well.

2.2. Cascaded regression

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 $(R^1, \ldots, R^t, \ldots, R^T)$ are successively applied starting from the initial shape estimate S^0 (Fig. 3). Given an image I, each regressor learns and estimates a shape increment δS and updates the face shape:

$$\delta S = R^t(I, S^{t-1}) \tag{2}$$

$$S^t = S^{t-1} + \delta S \tag{3}$$

where the tth regressor R^t updates the previous shape S^{t-1} to the new shape S^t [14]. 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 [22]. The method is called Cascaded Pose Regression (CPR) and was developed for general object alignment, including faces as well. The method owes its success to pose-indexed features where pixel positions used in the pixel difference features are stored relative to the object pose and are thus consistent across large pose variations. Random fern regressors were used at each stage of the cascade.

Cao et al. in their seminal work called Explicit Shape Regression (ESR) extend the idea from CPR [14]. Again, pixel difference features and fern regressors are used, however, the shape is jointly regressed as a vector which implicitly enforces a shape constraint in a non-parametric way (Eq. 2). They use a two-level boosted regression where each regressor in the cascade uses global features indexed relative to the nearest landmark (see Fig. 4). Each regressor in the cascade is also a cascade of primitive regressors (ferns) using fixed features. The authors used correlation-based feature selection when choosing the most discriminative features from the pool.

Kazemi and Sullivan in their work Ensemble of Regression Trees (ERT) improve upon the ESR method [23]. Instead of random ferns, gradient tree boosting ensembles are used. They also use shape-indexed features, indexed to the closest landmark, however, they transform the pixel positions in order to compensate for rotation, scale, and translation relative to the mean shape. A prior is introduced to favor closer pixel differences in their feature selection process. They use weights in the training-node split-error calculations in order to handle uncertain/occluded landmarks (the ground truth of some landmarks can be “turned off” when optimizing).

A somewhat different approach in the same framework is described in [24]. The algorithm is called Supervised Descent Method (SDM) and it presents shape-indexed features and cascaded regression as a Newton-type optimization of a non-linear least-squares problem. Basically, they use linear regression and local Scale-Invariant Feature Transform (SIFT) features from [25] on local patches centered on currently estimated landmark positions. Eq. (2) is then replaced by:

$$\delta S = W^t \phi^t(I, S^{t-1}) + b^t \tag{4}$$

where W^t and b^t are linear projection matrix and bias term respectively. ϕ^t is a non-linear global feature extraction function which

Table 1
Constrained regression methods summary.

Methods	Features	Regressor	Constraint	Year
BoRMaN [15]	Haar-like filters	SVR with Gaussian RBF	MRF	2010
Kazemi & Sullivan [16]	PHOG	Ridge	Tree structure	2011
SO-RF [18]	HOG & Gabor	Random forest	Structure graph	2012
Yang & Patras [19]	HOG & Gabor	Random forest	Sieves	2013
LEAR [20]	LBP	SVR	MRF sampling	2013

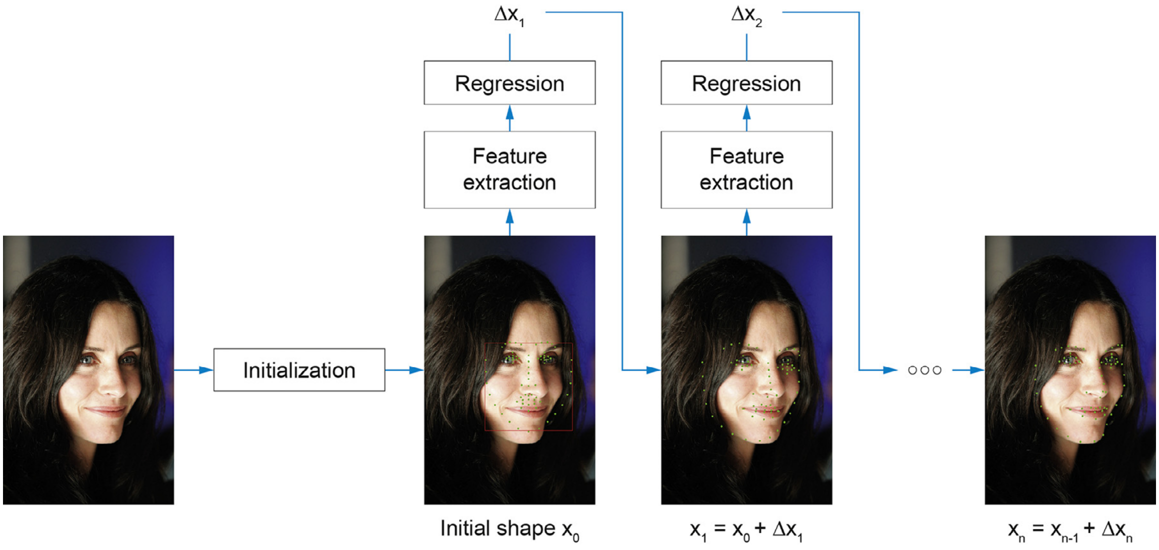


Fig. 3. Cascaded regression in a coarse to fine manner.

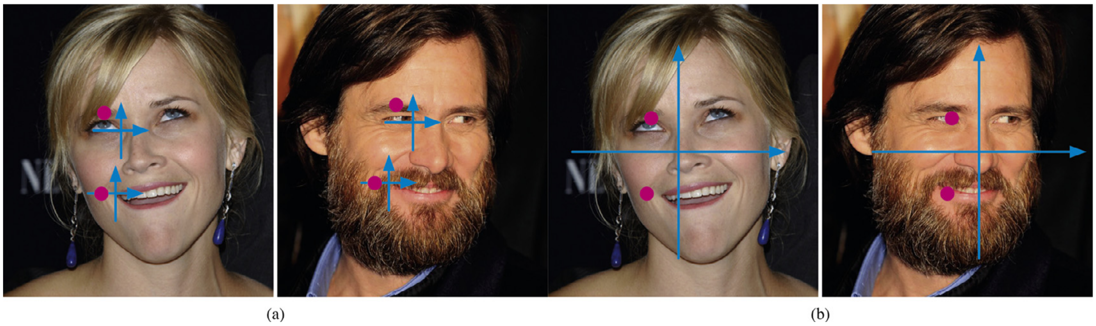


Fig. 4. Shape-indexed pixel positions (a) and globally indexed pixel positions (b). Shape-indexed features retain same semantic meaning regardless of face variations [14].

concatenates local features extracted around currently estimated landmarks.

A method called Local Binary Features (LBF) by Ren et al. [26] is an improvement of ESR [14] and SDM [24] methods. A random forest is used, trained to minimize the alignment error for individual landmarks to produce binary features. Local features are coded in a binary array by placing a one for leaves where the sample ends up while traversing the tree and zero otherwise. Features that are individually learned for each landmark are then concatenated into a global feature vector as input for ridge regression (linear regression with L_2 regularization). This method owes its success to a feature learning step where features are explicitly learned for the given custom task instead of manually crafted (such as SIFT). Due to the sparseness of the feature vectors, the inference phase can be reduced to traversing the forest and performing simple look-ups and additions. Ren et al. achieved a frame rate of 3000 FPS which is, of course, hardware dependent but impressive, nevertheless. In a later work, Luo et al. modified the forest to obtain probability features and used Probabilistic Random Forests (PRF) which modeled the probability for each sample belonging to a tree leaf node [27]. This slowed the algorithm down considerably (every sample must traverse every tree branch and the produced features are no longer sparse binary vectors), however, improved accuracy and stability (reduced noise during tracking on videos) is achieved. Another extension was presented in [28]. The main idea is to replace ridge regression with a neural network architecture utilizing a bottleneck layer. By doing this, the authors improved the accuracy, execution time, and reduced memory requirements of the original algorithm.

A similar, recent fast cascaded regression approach called Cascade Gaussian Process Regression Trees (cGPRT) was proposed in [29] by Lee et al. Features that are computed as difference-of-Gaussian filter responses on local retinal patterns referenced by the shape estimates are used instead of standard shape-indexed pixel differences as in ESR [14], ERT [23], and LBF [26]. For regression, Gaussian processes with a kernel modeled by trees are used, optimized for the individual landmarks. Both innovations improved the results from previous methods in [14,23,26].

Table 2
Cascaded regression methods summary.

Methods	Initialization	Features	Regressor	Year
CPR [22]	Random	Pixel difference	Random ferns	2010
ESR [14]	Random	Pixel difference	2-level ferns	2012
SDM [24]	Mean shape	SIFT	Linear	2013[a]
ERT [23]	Mean shape	Pixel difference	2-level trees	2014
LBF [26]	Mean shape	LBF	Ridge	2014
PRF [27]	Mean shape	PF	Ridge	2015
cGPRT [29]	Mean shape	DoG	GPRT	2015
MDM [30]	Mean shape	Local convolutions	RNN	2016[b]
LBF-NN [28]	Mean shape	LBF	Neural network	2018

[a] https://www.youtube.com/user/xiong828/videos.
[b] http://trigeorgis.com/mdm.

2.2.1. Recurrent cascaded regression

An interesting modification to the standard cascaded regression approach was recently proposed in [30]. The authors argue that there is a loss of knowledge between independently trained stages in the standard cascaded approach and propose a single Recurrent Neural Network (RNN) architecture that combines the training of all stages through the introduction of a state vector which serves as a mnemonic unit. The approach extends the classical Supervised Descent Method [24] with the use of an RNN, as already mentioned and, additionally, with local small CNNs as feature extractors instead of hand-crafted SIFT features. The authors conveniently named the method Mnemonic Descent Method (MDM) and showed that the introduced state vector partitions the training set into meaningful clusters with different descent directions in subsequent stages.

A similar approach was presented in [31]. Liu et al. use the same architecture, however, they investigate the correlation of neighboring landmarks in order to remove redundant information of overlapping patches. To this end, the mentioned correlation is explicitly modeled and utilized under a multi-task learning paradigm. Additionally, multi-scale images are used to enhance coarse-to-fine alignment through the use of an RNN.

2.2.2. Summary

A couple of major improvements were introduced with the adoption of the cascaded regression framework. The global shape information is no longer constructed by hand, it is implicitly deduced from the training set which demonstrated greater generalization ability. Furthermore, landmarks are regressed jointly, not individually, utilizing both local features and contextual information from neighboring landmarks. Finally, the complexity of the face alignment is broken down into a series of simpler problems through the cascaded architecture. Early stages of the cascade naturally focus on rough alignment dealing with head pose and shape variations while the later stages focus on local details and subtle variations in facial appearance.

In later developments, cascade stages are treated as a sequence of inputs which makes sense both from a practical and theoretical standpoint. A single model for all stages reduces memory requirements and retains knowledge between individual stages. Through the use of a state vector (“mnemonic”) inside the RNN architecture, the model can be made aware of decisions from previous stages and hence learn conditionally based on those decisions. As an additional bonus, this method can be naturally adjusted to tracking from video where information from previous frames can be efficiently utilized.

The summary of the described cascaded regression methods, with highlights of key differences, is presented in Table 2. These methods nearly saturate frontal and relatively constrained data sets. However, “in the wild” data sets are still challenging due to a weakness to initialization that these methods demonstrate. Local shape-indexed features, frequently used within the framework, do not capture a large enough context for samples initially far away from the ground truth. These situations occur frequently in unconstrained data sets including, e.g., full profile facial images. Strategies to address this issue within the framework will be described in the following sections.

2.3. Global feature initialization

Zhu et al. studied initialization strategies for face alignment and confirmed its importance for cascaded regression methods [32]. They managed to improve the robustness by initializing the ERT [23] cascaded regression method with an additional ERT [23] model trained on a subset of rigid landmarks to produce an improved initial shape. On the other hand, Valle et al. [33] addressed the problem of initialization sensitivity of the cascaded regression approach by introducing a CNN-based initialization stage. The method is called Deeply-initialized Coarse-to-Fine Ensemble (DCFE). A simple CNN architecture is used to estimate landmark probability maps using the whole face image as the input. These initial landmarks are then utilized as input to a 3D model fitting procedure which produces a robust and accurate initial face shape estimation. The fast ERT method [23] is used for precise alignment in the next stages of cascaded regression. The final stage uses separate models for facial regions in order to decouple the movements and achieve improved alignment for asymmetrical facial expressions.

A similar initialization approach was proposed in [34]. Again, the whole face image was used for feature extraction in the first stage. However, Kowalski et al. used K-cluster Regression Forests with Weighted Splitting (KRFWS) to regress the 3D head pose parameters which then served as initial landmark positions (after projection of mapped vertices to the image space). Subsequent stages are designed according to standard cascaded regression framework with KRFWS algorithm and shape-indexed Pyramid HOG features from [35].

An interesting method which combines deep convolutional networks for feature extraction and cascaded regression has been proposed in [36]. The authors named it Deep Cascaded Regression (DCR). It comprises of three modules. The first module is a convolutional/deconvolutional network which serves as a feature extractor for the other two modules. It produces a deconvolution layer of the same size as the input image. The second module performs initialization search. It uses the last deconvolution layer with a fully connected layer in order to learn, for each landmark separately, the probabilities of each pixel belonging to that particular landmark. It also generates a number of representative shapes from the training set using k-means. The landmark probabilities are then used to find the closest shape as the initialization. The third module performs cascaded regression using the initialization shape. Linear regression is used with a fully connected layer on the features ex-

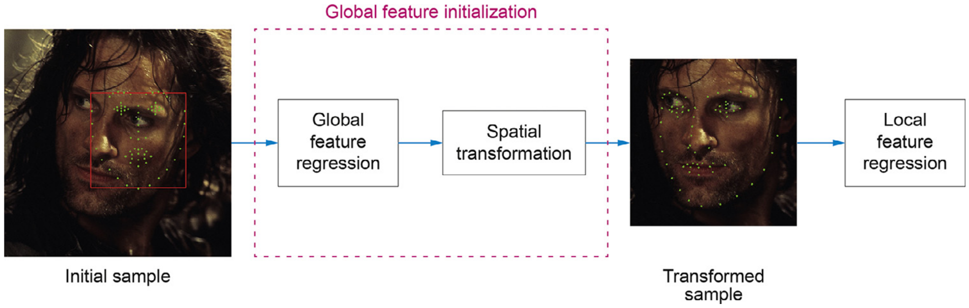


Fig. 5. Global feature initialization - the main idea is to use the whole face region to estimate the initial face shape in order to calculate and eliminate the spatial transformation with respect to the canonical shape.

Table 3
Global feature initialization summary.

Methods	Global features	Global regression	Spatial transform	Local cascaded regression	Year
DCR [36]	Convolutional	Encoder-decoder	Initialization search	Deconv. features + linear	2015
KRFWS [34]	PHOG	KRFWS	3D-APR	LBF [26] + KRFWS	2016
P-DSC-CR [37]	Convolutional	CNN	In-plane rotation	HOG + Lasso	2016
DCFE [33]	Convolutional	CNN	POSIT	ERT [23]	2018

tracted from the module-one deconvolution layer around the currently estimated landmark positions (shape-indexed) on fixed size patches.

A similar multiple module approach was proposed by Liu et al. in their work called Pose-insensitive Dual Sparse Constrained Cascade Regression (P-DSC-CR) [37]. They use a deep convolutional neural network to detect initial five landmarks and estimate the head pose. Separate cascaded regressors for each pose (frontal, profile) are learned. Cascaded regression is improved by adding dual sparse constraints. At each stage, landmark updates are first learned by Lasso regression which produces a sparse projection matrix. Then, the updated landmark positions are fitted to the sparse shape dictionary which produces the estimate for the current stage and is the input for the next stage. The dictionary is constructed using K-SVD [38] algorithm on training faces. Multi-scale Histogram of Oriented Gradients (HOG) [39] features centered on landmark positions are used for regression.

2.3.1. Summary

In general, this approach can be graphically summarized as shown on Fig. 5. Using an initialization stage with features from the whole face proved to be adequate to mitigate the initialization sensitivity problem of the cascaded regression approach. The initial stage takes a larger context around the initial mean shape as input which makes it robust to larger translation shifts from the ground truth face. Additionally, more complex algorithms with larger capacity (CNN) are usually used in the initial stage because of large input variance, while faster and more efficient features can be used in later stages to improve processing time while retaining high accuracy levels. The described methods are summarized in Table 3.

2.4. Cascade of experts

The greatest face shape variations come from different head poses with respect to the camera. A straightforward way to improve face alignment accuracy is to use multiple domain-expert models in parallel in order to make the cascaded regression approach more robust to various head poses.

A simple parallel cascaded regression approach was proposed in [40] by Feng et al. called Random Cascaded Regression Copse (R-CR-C). Three parallel cascaded regression threads are trained on random subsets of the training set and used at inference phase. The regressors are plain ridge regressors using the Sparse Auto-Encoder features. A very similar approach was proposed in [41] using the FEC-CNN architecture [42] (described in Section 2.5.1) as the backbone. The method achieved very good results on the recent Menpo Challenge [43].

A more elaborate method was proposed by Xiong and De la Torre in their work Global Supervised Descent Method (GSDM) where they extend the original SDM [24] method to handle large pose variations [44]. The problem is again cast in a non-linear optimization framework where the aim is to find a globally better minimum by partitioning the domain of the optimization into regions of similar gradient descent. Mathematical theory is demonstrated that ensures such partitions exist, and a procedure is given on how to find them. A separate SDM [24] model is trained for each region. The solution was applied to face tracking in videos.

Zhu et al. in their work Cascaded Compositional Learning (CCL) [45] developed a similar idea. Again, the optimization space is divided into multiple domains of homogeneous descent and separate experts trained for each domain. However, Zhu et al. added an explicit module to handle the initial domain selection instead of relying on the previous frame output as in the GSDM method. It makes the method more effective on images. The outputs of the individual experts are combined in a learning framework that directly optimizes landmark positions.

Dong et al., Zhu et al., and Rampal et al. have similar ideas on how to handle occlusions and extreme poses in their respective works Robust Discriminative Regression (RDR), Ensemble of Model Recommendation Trees (EMRT), and Ranked Parts Based Models (RPBM) [46–48]. The idea is to train multiple cascaded regressors using a heuristically determined subset of landmarks when extracting shape-indexed features. They all used linear regression and SIFT/HOG features. The difference is in how their output is combined at each stage. Dong et al. and Zhu et al. both use recommendation trees to learn weights used for the linear combination of estimates (quadratic programming is used to find the optimal weights at each node). Rampal et al. train a Support Vector Machine (SVM) to produce a ranking for each model using shape-indexed HOG features.

Finally, a probabilistic approach was proposed by Zhu et al. called Coarse-to-Fine Shape Searching (CFSS) method [49]. The main contribution is to search a shape sub-space at each stage in a coarse-to-fine manner from which initial shapes are sampled for regression. First, a shape library is created using Procrustes analy-

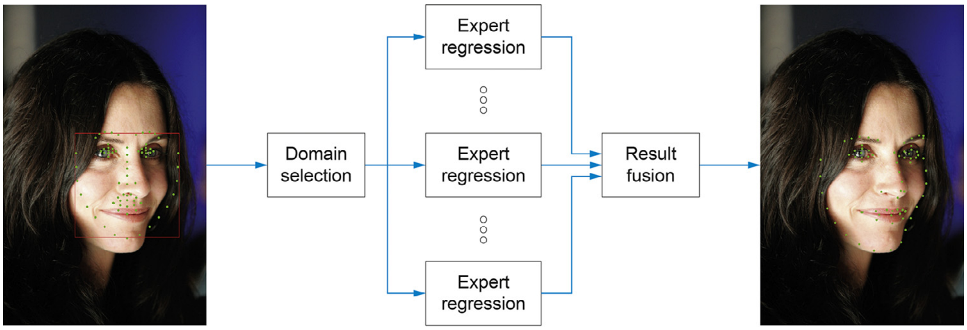


Fig. 6. The cascade of experts can be divided into three elements: domain selection, expert regression, and result fusion.

Table 4
Cascade of experts summary.

Methods	Domain selection	Expert regression	Result fusion	Year
R-CR-C [40]	Random	Sparse Auto-Encoder + Ridge	Average	2015
GSDM [44]	Homogeneous descent	SDM [24]	Single result	2015[a]
CCL [45]	Homogeneous descent	LBF [26]	Composition ridge regression	2016[b]
RDR [46]	Facial region	SIFT + linear	Learned weighted average	2015
EMRT [47]	Head pose & occlusion	SIFT + SVM	EMRT	2015
RPBM [48]	Facial region	HOG + linear	SVM ranking	2015
CFSS [49]	Shape sub-space distribution	BRIEF/SIFT + linear	Single result	2015[c]

[a] http://goo.gl/EGiUFV.
[b] http://mmlab.ie.cuhk.edu.hk/projects/compositional.html.
[c] http://mmlab.ie.cuhk.edu.hk/projects/CFSS.html.

sis. At each stage the goal is to find a finer shape sub-space represented by a sub-space center and a more narrow normal shape probability distribution. The initial probability distribution is uniform representing equal chances for every shape in the library to be selected. Using the estimated shape probability distribution, a number of initial shapes are sampled and the regression performed for each. The outputs from the regressors are combined using weights obtained through the dominant set approach and form the sub-space center. Using the sub-space center, the probability distribution for the next stage is also estimated. At the last stage, the sub-space center is the final estimate. At each stage, a cascade of linear regressors using either BRIEF [50] or SIFT [25] features (accuracy vs speed trade-off) is used.

2.4.1. Summary

Cascade-of-experts is a logical approach to reducing complexity by dividing the problem into sub-domains as illustrated on Fig. 6. However, it comes with a greater computational cost since multiple models are trained and then used at inference phase. The accuracy boost is evident in the respective papers but often comes with a cost of doubling or even tripling inference time and memory requirements. It makes the approach impractical in many scenarios. The summary of the described methods is presented in Table 4.

2.5. Deep learning

Deep learning methods have recently gained in popularity due to the advances both in hardware and optimization techniques. They have been applied in many computer vision fields including face alignment as well.

However, there has not been much success in taking a single deep CNN architecture and training it to accurately locate landmarks on a face image. One of the reasons is the need for large data sets in order to make such an approach successful. Wu et al. proposed to unify data sets with different annotations to increase both the size and the variance of the training set [51]. An architecture called Deep Variation Leveraging Network (DVLN) was used, consisting of two CNN networks: Dataset-Across Network (DA-Net) and Candidate-Decision Network (CD-Net). The DA-Net was trained on the unified training set where the deep layers of the network were shared across sub-sets with different annotations while the last fully connected layers were specific for each annotation configuration. Additionally, they normalized the data sets so that a single profile view is present reducing the complexity of the problem. The CD-Net was trained to recognize the view of the facial image (left or right profile) and select the correct output of the DA-Net which takes as input normal and flipped images. The method achieved the second best result on the Menpo Challenge [43].

Another direct regression approach using deep learning was proposed in [52] where a doubly CNN architecture [53] was used which is computationally more efficient than regular convolutions along with Fourier feature pooling to build strong holistic representations. In order to encode landmark correlation, the authors designed a layer with linear low-rank learning instead of fully connected layer as the output layer.

An interesting idea was explored by Shao et al. using a deep learning model named Multi-Center Network (MCNet) [54]. A CNN architecture based on VGGNet was trained in a standard way for face alignment, however, the authors used that pre-trained model and its shared deep features to separately fine-tune seven landmark regions improving the precision of the original model.

2.5.1. Deep cascaded regression

Since cascaded regression achieved breakthrough results for face alignment, the logical next step was to combine it with deep learning. Sun et al. were pioneers in this area with their work called Deep Convolutional Network Cascade (DCNC) and proposed a cascaded regression approach with three stages of convolutional networks [55]. A shape with only five landmarks is estimated. At each stage, predictions from multiple networks are fused together in order to improve the accuracy and reliability of the estimation. The first stage networks take the whole face image as input and predict initial estimates of the landmark positions. The next two stages use patches centered on the estimated landmark positions as input from the previous stage and refine the estimations to achieve higher accuracy.

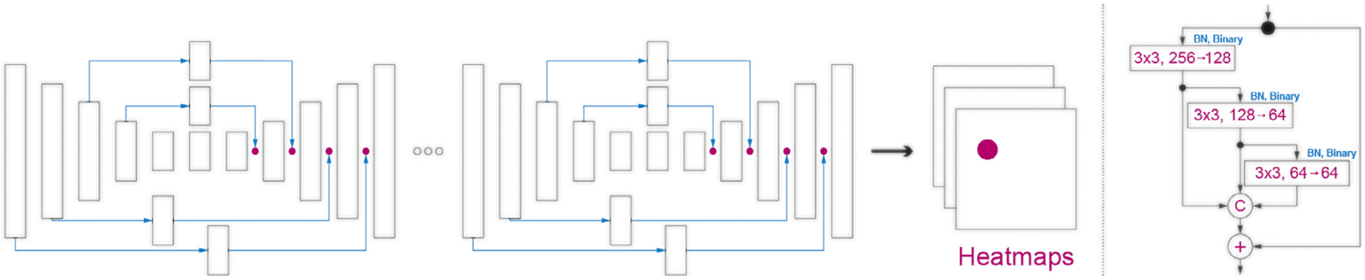


Fig. 7. The hourglass CNN architecture and residual block used in [62].

Table 5
Deep learning methods summary.

Methods	Architecture	Positions	Heatmaps	Year
CF-CNC [56]	Cascaded CNNs	Yes	No	2013
CFAN [57]	Cascaded auto-encoders	Yes	No	2014
DVLN [51]	Two coupled CNNs	Yes	No	2017
MCNet [54]	Single CNN	Yes	No	2017[a]
DCNC [31]	Cascaded local & shared CNNs	Yes	No	2017
DAN [58]	Cascaded CNNs	Yes	No	2017[b]
FEC-CNC [42]	Cascaded local CNNs	Yes	No	2017
FAN [62]	Stacked HGs	No	Yes	2017[c]
DSRN [52]	Single CNN	Yes	No	2018[d]
LAB [65]	Stacked HGs	No	Yes	2018[e]
DeCaFA [61]	Cascaded U-nets	Yes	Yes	2019
AWing [68]	Stacked HGs	No	Yes	2019[f]
Liu et al. [66]	Stacked HGs	No	Yes	2019
KDN [67]	Stacked HGs	No	Yes	2019
HRNet [69]	Parallel high-to-low CNNs	No	Yes	2020[g]

[a] https://github.com/ZhiwenShao/MCNet.
[b] https://github.com/MarekKowalski/DeepAlignmentNetwork.
[c] https://www.adrianbulat.com/face-alignment.
[d] https://github.com/xinxinmiao/DSRN.
[e] https://wywu.github.io/projects/LAB/LAB.html.
[f] https://github.com/protossw512/AdaptiveWingLoss.
[g] https://github.com/HRNet/HRNet-Facial-Landmark-Detection.

In a similar work called Coarse-to-Fine Convolutional Network Cascade (CF-CNC), Zhou et al. proposed to separate the detection process for inner and contour points [56]. The first stage neural network estimates the bounding boxes for inner and contour points separately. The second stage gives an initial holistic prediction of inner and contour points, also separately. The third stage refines the six facial parts, computes their rotations and normalizes them before giving the patches to the fourth and final stage which makes final refinements. The contour points do not utilize the third and fourth stages. Zhang et al. use a similar framework with stacked auto-encoder networks in their work named Coarse-to-Fine Auto-encoder Networks (CFAN) [57].

Kowalski et al. combined deep learning networks with the cascaded regression framework in [58] naming the method Deep Alignment Network (DAN). All stages in the cascade use global facial region as input to the deep convolutional networks. In order to keep the advantages of shape-indexed features and transfer of knowledge between stages, the authors implemented connection layers that generate inputs for the next stage based on the output of the current stage. According to their experiments, two stages were enough to achieve convergence. A similar two-stage framework was used in [59], however, the authors investigated the optimal loss for landmark regression arguing that the universally used L2 loss is too sensitive to outliers in the data set. Through intuition and experiments, they derived a formulation of the Wing loss which is able to balance the influence of both small and large errors during training. This enabled them to use relatively simple CNN architectures in both stages while achieving competitive accuracy.

All of the above-mentioned methods had their stages in the cascade trained individually and separately from the others. In [42], the authors claim that the cascade could benefit from joint training of the stages enabling the flow of information between them. A Fully End-to-End Cascaded CNN (FEC-CNN) architecture is introduced which uses local shape-indexed CNN features in each stage and is optimized jointly using Stochastic Gradient Descent (SGD) and back-propagation. The biggest challenge was to generate gradients of shape-indexed patches of the image with respect to the input shape from the previous stage. They managed to successfully formulate the derivations by drawing inspiration from Spatial Transformer Networks [60]. The experimental section confirmed the benefits of the end-to-end training procedure. In a later work, Dapogny et al. strove for a similar goal of end-to-end optimization in their work Deep Convolutional Cascade for Face Alignment (DeCaFA) [61]. This was achieved, however, using fully convolutional stages with U-net architecture and transfer layers designed to produce landmark-wise attention maps. In order to use heterogeneous data (multiple data sets with different annotations), they used chaining of multiple transfer layers ordered by the density of landmarks in the face shape.

2.5.2. *Heatmap regression*

A recent method in [62] reached a saturation performance by using a modern deep CNN architecture and a generated large data set making the face alignment problem solved in most scenarios. The authors used a state-of-the-art hourglass CNN architecture with a novel residual block (see Fig. 7) and trained each landmark's location as a heatmap which produced estimates on position certainty as well. The same network architecture was trained to convert 2D landmark annotations to 3D and used to create a large-scale 3D facial landmark data set with approximately 230k images. A similar approach was proposed in [63] which achieved state-of-the-art results in the recent Menpo Challenge [43]. The authors added a supervised face transformation step which eliminates rigid face transformations based on the output of a face detector and its reduced subset of detected landmarks [64]. This improved the robustness of the method and reduced the complexity of the problem for the stacked hourglass CNN training.

The success of both methods inspired other researchers to investigate the heatmap regression approach in greater detail using the same stacked hourglass architecture [65–68]. An interesting improvement was proposed by Wu et al. in their work named Look at Boundary (LAB) [65]. The main premise is that most landmarks in the face shape are ill-defined even in frontal pose, thus, they introduce face boundaries as a more suitable face geometry representation. The stacked hourglass architecture is, therefore, used to estimate high quality boundary heatmaps using adversarial learning. The boundary heatmaps are then driving the regression CNN to produce accurate landmark positions. An additional benefit of the boundary paradigm is the innate ability to represent heterogeneous annotations enabling the architecture to use a large unified data set for training. Instead of using an additional CNN to produce landmark positions, Wang et al. estimate both landmark and boundary heatmaps using the stacked hourglass architecture [68]. However, their main contribution is modifying the Wing loss introduced in [59] and applying it to heatmap regression. Their Adaptive Wing (AWing) loss is designed to be more sensitive to small errors on foreground and less on background pixels confirming the inferiority of the L2 loss once again.

Following similar reasoning, Liu et al. also stress semantic ambiguity of contour landmarks [66]. Instead of introducing boundaries, they opted for a probabilistic model of the “real” ground-truth. Landmark updates during training iterations are used to discern between random movements due to annotation noise and meaningful movements towards ground-truth. The probabilistically modeled “real” ground truth is then used in later iterations to achieve stable and more accurate convergence. Chen et al. addressed the same problem using Kernel Density Deep Neural Network (KDN) [67]. Instead of assuming a Gaussian distribution of the heatmap regression, their model can estimate a more general probability distribution, e.g. multimodal or asymmetric distribution. Furthermore, they extend the stacked hourglass architecture inspired by cascaded regression in order to propagate the estimated probability distribution between stages.

While all of the above methods aim to produce a low resolution representation from which the landmarks are predicted, Wang et al. argue that high resolution representation is beneficial for all spatial vision tasks [69]. In their novel HRNet architecture, parallel high-to-low convolutions are employed with a multi-resolution fusion scheme to exchange information across resolutions. The validity of their hypothesis and superiority of their architecture design is demonstrated on a wide range of vision problems: human pose estimation, semantic segmentation, object detection, and face alignment.

2.5.3. Summary

Judging by the lack of successful research with a single, simple deep learning model for face alignment, it seems that the problem is too complex and data sets too small for such a straightforward approach. Thus, complex architectural and training procedures need to be implemented to achieve competitive results. The combination of deep learning models and cascaded architecture where the problem is broken down into manageable sub-problems is a promising solution. Similarly to the group of methods from Section 2.3, the first stage in the deep regression cascade uses the whole face region to predict a subset of landmark positions. However, later stages also utilize CNN architectures with different techniques to focus the network on finer details. Methods from this group achieve great results “in the wild” though it comes with a greater computational cost since demanding CNN architectures are used throughout the cascade. Real-time performance is possible with the use of a modern Graphical Processing Unit (GPU).

Another promising deep learning direction is the use of fully convolutional networks for heatmap regression with a widespread adoption of the stacked hourglass architecture. These methods achieve state-of-the-art results across different benchmarks, however, real-time performance is not possible even with a high-end GPU. This is understandable since an additional decoder block is necessary in the CNN architecture to estimate each pixel heatmap value. The summary of the described methods is presented in Table 5.

3. 3D Face alignment

Due to the ambiguity and self-occlusion of 2D landmarks in more extreme poses, 3D landmark alignment has gained traction in recent years. The ambiguity is most notable on the contour landmarks in semi-frontal and profile poses as can be seen on Fig. 8. 3D landmarks maintain physical meaning of the contour across the whole range of head poses while 2D landmarks change semantics which introduces additional complexity in the training process.

Several different approaches have emerged seeking to exploit the coherence of 3D facial structure mostly differing in representation of the regression target. However, any 3D alignment approach needs reliable 3D annotations along with the images. One way of obtaining the necessary “ground-truth” information is by utilizing specialized 3D imaging hardware which produces a 3D point cloud corresponding to the pixels in the image [70–72]. These raw results can not be used directly because each facial scan has different topology of vertices and must be registered under a single mesh topology. This is usually done by employing an Iterative Closest Point (ICP) algorithm and its variants [73,74]. Due to the complicated acquisition process, these data sets are usually collected in a controlled environment with relatively few subjects.

A different approach to building a 3D alignment data set is to fit a 3D Morphable Model [9,75,76] on existing large 2D data sets. 3DMM is a statistical model of the face shape built from a data set of registered facial scans. Since it is a vital part across the whole 3D face alignment pipeline, we will introduce the general concept of 3DMM construction and representation.

We can define the 3D face shape (mesh) of N vertices as a $3N \times 1$ vector of their 3D coordinates:

$$\mathbf{S}_{3D} = [x_1, y_1, z_1, \ldots, x_N, y_N, z_N]^T \tag{5}$$

Using Principle Component Analysis (PCA) and its variants on a data set of registered 3D shapes, a 3DMM can be constructed and defined in the following way:

$$\mathbf{S}_{3D}^N = \bar{\mathbf{S}}_{3D} + \sum_k^{N_I} p_k^I \mathbf{S}_k^I + \sum_k^{N_E} p_k^E \mathbf{S}_k^E \tag{6}$$

where $\bar{\mathbf{S}}_{3D}$, $\mathbf{S}^I$, and $\mathbf{S}^E$ represent the mean shape, identity or face structure bases, and expression or action bases, respectively (see Fig. 9). The corresponding parameters which control their linear combination in the 3DMM are represented by $\mathbf{p^I} = [p_1^I, \ldots, p_{N_I}^I]$ and $\mathbf{p^E} = [p_1^E, \ldots, p_{N_E}^E]$. They are often combined in a single parameter vector $\mathbf{p} = [\mathbf{p^I}, \mathbf{p^E}]$ controlling the non-rigid transformations of the 3DMM:

$$\mathbf{S}_{3D}^N = \mathcal{N}(\mathbf{p}) \tag{7}$$

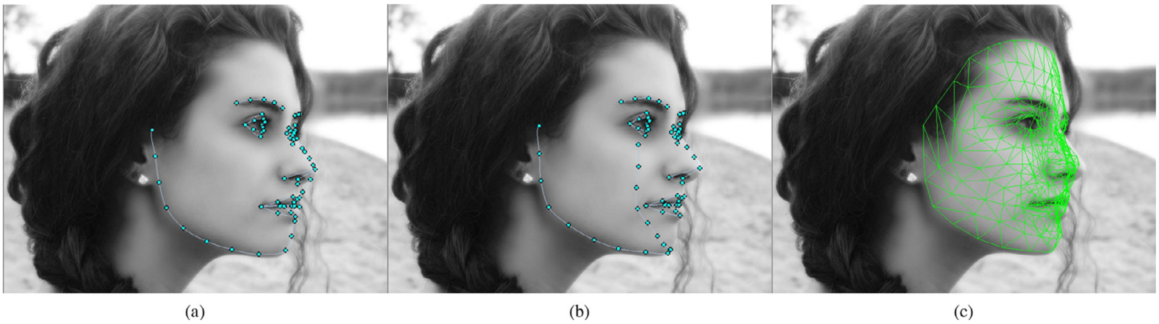


Fig. 8. Different regression targets: standard sparse 2D landmarks (a), sparse 3D landmarks (b), and dense 3D landmarks (c). The largest difference between 2D and 3D landmarks can be observed on the contour landmarks.

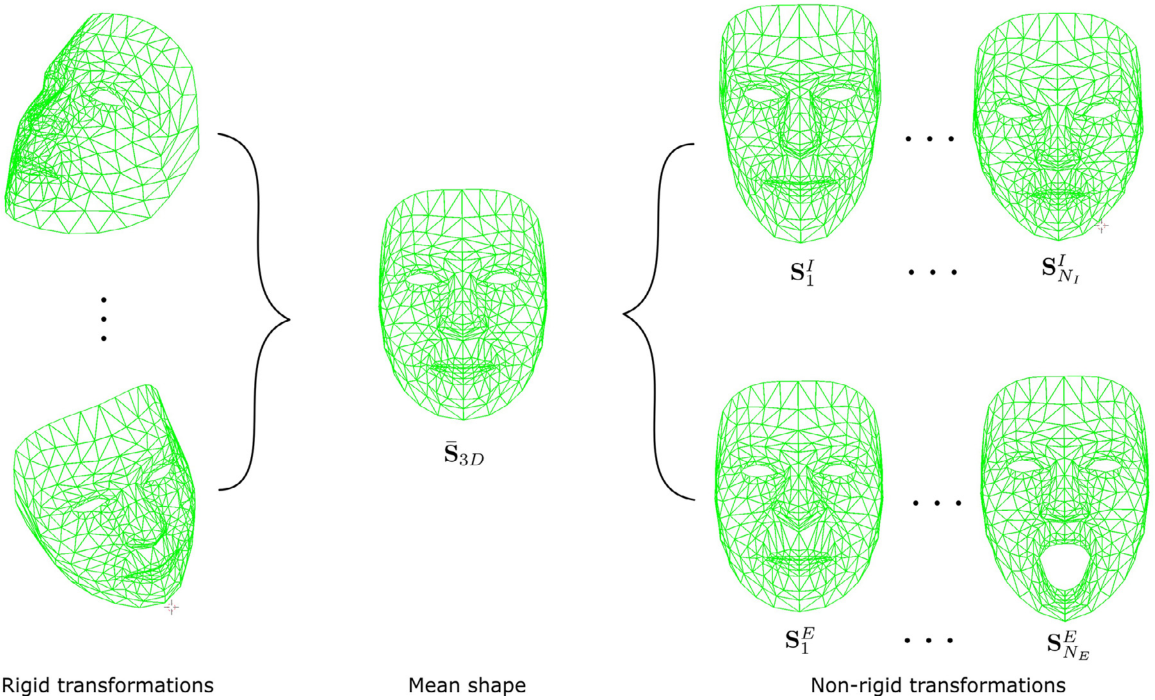


Fig. 9. Components of the 3DMM.

The resulting 3D mesh $\mathbf{S}_{3D}^N$ is in a normalized shape space. In order to bring it to the 2D space on the image plane, a model of the camera with its transformations needs to be included. Weak perspective projection of the pinhole camera model is usually employed with six degrees of freedom (scale, three rotations, and two translations) which can be represented by a vector $\mathbf{c} = [s, r_x, r_y, r_z, t_x, t_y]$. The projection can then be defined as:

$$\mathbf{U} = \mathcal{W}(\mathbf{p}, \mathbf{c}) \equiv \mathcal{P}(\mathcal{N}(\mathbf{p}), \mathbf{c}) \tag{8}$$

where $\mathbf{U} = [x_1^u, y_1^u, \ldots, x_N^u, y_N^u]$ represents the projected 2D coordinates in the image.

Finally, a subset of projected 2D coordinates $\mathbf{U_L}$ corresponding to the annotated 2D landmarks in the image can be used to drive the optimization process to determine the "ground-truth" 3D annotations:

$$\underset{(\mathbf{p},\mathbf{c})}{\arg\min} \|\mathcal{W}_L(\mathbf{p}, \mathbf{c}) - \mathbf{S}^*\|^2 \tag{9}$$

The goal of the optimization process is to find the rigid ($\mathbf{c}$) and non-rigid ($\mathbf{p}$) transformation parameters of the 3DMM that minimize the distance between the annotated 2D landmarks and their corresponding 3D shape projections. In order to mitigate the inconsistency of the 2D landmarks, dynamic correspondence of the contour landmarks is employed such as landmark marching [77]. This optimization can be performed on large and diverse 2D alignment data sets to produce an "in-the-wild" 3D alignment data set under consistent annotations with minimal supervision [62].

We classified the existing approaches by the regression target, and consequently their output, into three categories. Sparse 3D alignment methods optimize the subset of 3D mesh vertices or their projections corresponding to the usual 2D mark-up. 3DMM alignment methods optimize the 3DMM transformation parameters $\mathbf{p}$ and $\mathbf{c}$ and can produce a dense 3D shape by applying the transformations. Finally, direct 3D alignment methods do not utilize a 3DMM but can still produce a dense 3D shape through creative output and optimization design. Each group of methods will be covered more thoroughly in the following sections.

3.1. Sparse 3D alignment

Since 2D alignment was extensively researched in the last 20 years, the straightforward approach is to simply replace the 2D annotations with the corresponding 3D annotations or their projections [78–80]. As already mentioned in Section 2.5, Bulat and Tzimiropoulos experimented with both 2D and 3D alignment using the same CNN architecture (named 2D-FAN and 3D-FAN) where they observed a slight improvement in accuracy by switching to 3D alignment [62]. The model 3D-FAN was trained on the 300W-LP-3D [81] data set which is constructed by 3DMM fitting on the 300W data set.

In a recent work, Deng et al. use both 2D annotations and the corresponding 3D projections to drive the optimization process in a cascaded framework [82]. Hourglass CNN architectures are used in two stages. The first stage is used for coarse alignment and jointly estimates 2D landmarks in both frontal and profile poses exploiting the correspondence of different annotations. The second stage optimizes the corresponding projected sparse 3D landmarks to refine the prediction and provide full-pose alignment results.

One of the earliest attempts at 3D alignment was proposed by Jourabloo and Liu in their work called Pose-Invariant Face Alignment (PIFA) [83]. As the name suggests, they were motivated to achieve face alignment covering even full profile poses. This method is interesting because it can be seen as a predecessor for 3DMM alignment methods. A model is constructed from 3D facial scans as well and used to produce 3D annotations on existing 2D alignment data sets. However, the 3D Point Distribution Model (3DPDM) is constructed using the sparse set of landmarks and can not produce a dense 3D mesh. A standard cascaded regression approach is used where at each stage optimization alternates between rigid and non-rigid transformation parameters.

3.2. 3DMM Alignment

In a later work, the same authors extended their approach by utilizing a 3DMM to produce a dense 3D output and replaced hand-crafted features with CNNs [84]. With an additional introduction of 3D aware inputs for CNN, they managed to improve their previous results. A similar approach was presented in [81] by Zhu et al. named 3D Dense Face Alignment (3DDFA). The cascaded regression approach was used to iteratively update the 3DMM transformation parameters with CNNs as stage regressors. In addition to the standard RGB input to the CNN, a novel channel called Projected Normalized Coordinate Code (PNCC), specifically designed to transfer the output of previous stages of the cascade, was used as input as well. A very important additional contribution is a method to expand the 300W training set with 3D annotations and augmenting it with generated profile samples to create the 300W-LP data set. This augmentation technique will be covered in more detail in Section 6.1.1.

In an attempt to connect the stages of the cascaded regression framework in an end-to-end manner, a technique already proven beneficial for 2D alignment, Jourabloo et al. introduced a visualization layer between the stages in their recent work [85]. The differentiable visualization layer generates an image of a 3D face using surface normals based on currently estimated 3DMM parameters to be used as an additional input to the next CNN stage. This allowed the flow of information between the deep learning stages and an end-to-end optimization leading to faster training convergence.

Liu et al. focused on the supervision signal and loss function of the 3D alignment training in their work Dense Face Alignment (DeFA) [86]. In addition to the standard sparse supervision (distance between 2D annotations and projections of the corresponding vertices), two additional terms to the loss function were added. The first term includes SIFT matching of vertices on pairs of images of the same face which enables dense alignment. The other term includes visible contour supervision using Holistically-nested Edge Detection (HED) [87] as the “ground-truth”. Additional supervision signals allowed them to achieve high accuracy using a single CNN without cascading.

Finally, Bhagavatula et al. emphasized the limitations of the 3DMM and its flexibility to model unseen faces [88]. Their approach is based on a 3D Spatial Transformer Network (3DSTN) which estimates the camera projection matrix as usual and parameters of the Thin Plate Spline (TPS) [89] warping function which performs the non-rigid 3D shape transformation. Using a non-linear warping function eliminated the indirect need of large data sets of 3D facial scans required by the 3DMM. This method shares the same goal as the next group of methods and can serve as a good transitional example.

3.3. Direct 3D alignment

The latest direction for 3D face alignment is to skip the 3DMM and its constraints to directly regress a dense 3D face shape. The straightforward approach of simply using a fully connected layer with an output for each vertex coordinate is not feasible due to the large number of vertices. Such a layer would be impractically large and challenging to train.

Nevertheless, Jackson et al. recently proposed the first direct 3D alignment method called Volumetric Regression Network (VRN) which uses a volumetric representation of the 3D face shape [90]. Such a representation allows them to use a fully convolutional network architecture and convert the problem into a 3D binary volume segmentation. The 3D face shape is discretized into voxels, a 3D binary volume, where the label of the voxel represents if it belongs to the face or the background. Two stacked hourglass CNNs are used with RGB image as input and binary volume as output. Another example of a direct representation was presented by Yu et al. using a per pixel 2D flow between the input image and the synthetically rendered image of a 3DMM [91]. An encoder-decoder architecture was trained on both synthetic and real examples (300-W-LP).

In a later work, a more efficient direct representation was proposed by Feng et al. named Position Map Regression Network (PR-Net) [92]. A UV position map is used wherein the RGB values for each UV texture coordinate of the 3D face model are replaced by the 3D coordinates of the vertices. This again allows a fully convolutional architecture, however, it is more efficient than the volumetric representation [90] which needs to discretize the interior of the head as well which is redundant for the face alignment problem. Such an architecture allowed them to use a single lightweight CNN architecture and achieve superior accuracy.

3.4. Summary

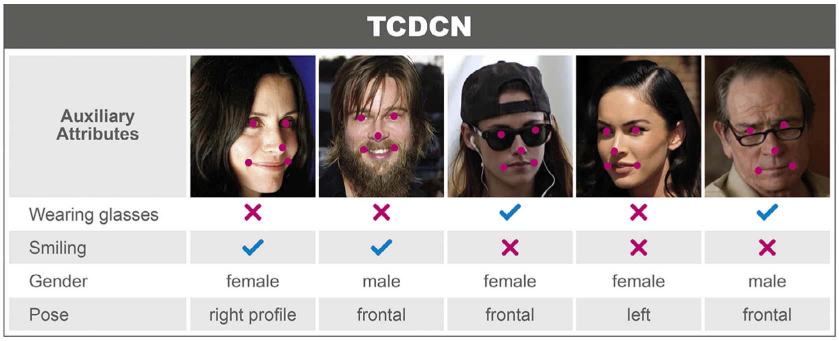
3D face alignment is a necessary step in the right direction if we want to achieve robustness across the full range of head poses since manual annotation of self-occluded landmarks is not feasible. The sparse 3D alignment methods benefit directly from the consistent and complete annotations even in full profile poses. In order to efficiently achieve dense alignment producing a detailed 3D facial mesh, 3DMM alignment methods optimize the 3DMM rigid and non-rigid parameters. Finally, direct 3D alignment methods eliminate the constraints of the 3DMM and directly optimize the dense 3D shape representation. All of these methods achieve full pose face alignment that 2D alignment methods can not achieve by design.

The biggest obstacle, however, for a wider adoption of this approach is the lack of annotated large-scale 3D data sets. The 3DMM is constructed from a data set of 3D facial scans using PCA which means that the flexibility of the model directly depends on the variety of the samples. The current publicly available data sets with 3D facial scans are collected on a scale of a hundred subjects. Collecting such a data set is cumbersome and expensive because of the additional hardware requirements. On the other hand, 2D alignment data sets contain “in-the-wild” images of thousands of subjects. Automatic re-annotation of these data sets using 3DMM

Table 6
3D face alignment summary.

Methods	Architecture	Target	Year
Tulyakov & Sebe [78]	Cascaded regression	Sparse 3D shape	2015
Gou et al. [79]	Cascaded regression	Sparse 2D shape + 3DMM	2015
PIFA [83]	Cascaded regression	3DPDM	2015
Zhao et al. [80]	Single CNN	Sparse 3D shape	2016
3DDFA [81]	Single CNN	3DMM	2016[a]
3D-FAN [62]	Stacked HGs	Sparse 3D shape	2017[b]
PAWF [84]	Deep cascaded regression	3DMM	2017
Jourabloo et al. [85]	Deep cascaded regression	3DMM	2017
DeFA [86]	Single CNN	3DMM	2017[c]
3DSTN [88]	Single CNN	3DMM + TPS	2017
VRN [90]	Stacked HGs	3D volume	2017[d]
Yu et al. [91]	Encoder-decoder	2D flow	2017
CMHM [82]	Cascaded HGs	Sparse 2D/3D shape	2018
PRNet [92]	Encoder-decoder	UV position map	2018[e]

[a] https://github.com/cleardusk/3DDFA.
[b] https://www.adrianbulat.com/face-alignment.
[c] http://cvlab.cse.msu.edu/project-pifa.html.
[d] http://aaronsplace.co.uk/papers/jackson2017recon/.
[e] https://github.com/YadiraF/PRNet.



Wearing glasses	✗	✗	✓	✗	✓
Smiling	✓	✓	✗	✗	✗
Gender	female	male	female	female	male
Pose	right profile	frontal	frontal	left	frontal

Fig. 10. Some of the attributes used in TCDCN [94].

fitting is a viable alternative. However, even though direct 3D alignment methods eliminate the explicit constraint of the 3DMM, it is still there implicitly through the construction of the training sets.

Nevertheless, recent years have seen an advancement in depth cameras making them smaller and cheaper to the point of integrating such cameras into mobile phones. With such advancements, the barriers for large-scale data set collection are becoming smaller making this approach viable in the future. The summary of the described methods is presented in Table 6.

4. Multi-task learning

Multi-task learning has proven to be effective in many research areas [93]. One of the first attempts for face alignment was proposed by Zhang et al. in their work named Tasks-Constrained Deep Convolutional Network (TCDCN) [94]. The main idea is to jointly learn auxiliary attributes along with landmark detection (Fig. 10). They proved that by learning the auxiliary relevant attributes, the complexity of the shape detection problem could be reduced. However, because of the different task complexities and convergence rates, modifications to simple multi-task learning algorithms were needed, and inter-task correlation modeling was introduced in the objective function via the covariance matrix to improve the performance and to analyze relations between attributes and landmarks. Additionally, a dynamic task coefficient was introduced to address the problem of different convergence rates between tasks. By using this technique, the learning process of some tasks could be turned off or the impact in the objective function reduced, if needed. 22 different attributes were used and annotated in the training set.

Another early attempt was presented by Zhao et al. by modeling and exploiting the relationships between multiple face analysis tasks (head pose, facial expression, and landmark detection) for mutual benefit. This unified method is called iterative Multi-Output Random Forests (iMORF) [95]. Random patches of the face image are used similarly to CRF [96], and a hybrid cost function is optimized which models the quality of each task using associated weights. The weights are dynamically adapted in order to put more emphasis on the head pose at the top nodes until its quality achieves sufficient purity of classification. Afterwards, facial expressions take precedence, again until sufficient purity is reached. Lastly, the landmark regression is performed. In the next phase, cascaded regression is employed to further refine the landmark positions (learn the update vectors) along with head pose and facial expression classification. In addition to the shape-indexed appearance features, shape related features are added that are modeled as distances and ratios of the landmark positions.

Face alignment or facial landmark detection highly depends on face detection which makes it logical to combine these two problems under a multi-task learning framework. Chen et al. in [97] demonstrated that alignment helps detection and managed to obtain improved results using joint learning in a standard cascaded regression framework using boosted regression trees as in [14].

Later on, the same idea was examined in [98] using CNNs. Again, cascaded regression framework is used with three stages of CNNs where each stage has a different goal under a paradigm of coarse-to-fine refinement. These tasks are performed sequentially in the cascade: face region proposal, face bounding box refinement, and face alignment with five landmarks. Additionally, each stage can reject the region as a non-face meaning it simultane-

Table 7
Multi-task learning methods summary.

Methods	Architecture	Additional tasks	Year
TCDCN [94]	Single CNN	Auxiliary attributes	2014[a]
iMORF [95]	Cascaded regression	Head pose + facial expression	2014
Chen et al. [97]	Cascaded regression	Face detection	2014
Zhang et al. [98]	Deep cascaded regression	Face detection	2016
MHM [99]	Cascaded HGs	Face detection	2019
Zhao et al. [100]	Encoder-decoder	Face segmentation	2019

[a] http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html.

ously performs face classification. This work was subsequently extended by an additional stage using a Multi-view Hourglass Model (MHM) [99] to produce a dense set of facial landmarks by exploiting the correspondence of semi-frontal and profile annotations of the Menpo data set.

An interesting combination of face alignment and segmentation was proposed by Zhao et al. in their multi-task learning work [100]. An encoder-decoder CNN architecture is used where the encoder is conditioned for the face alignment task and the decoder estimates the segmentation mask. A boost in accuracy is achieved for both of these highly correlated tasks through joint optimization.

4.1. Summary

It is evident from described methods that face alignment can benefit from multi-task learning with related tasks such as: face detection, head pose estimation, expression classification, gender estimation etc. This is especially useful in deep learning and cascaded regression frameworks where more general facial features can be learned and shared across related problems. Additionally, knowledge between tasks is often complementary and can be used to boost accuracy (e.g. smiling expression and landmark detection surrounding mouth region). The summary of the described methods is presented in Table 7.

5. Occlusion modeling

Faces are often occluded in unconstrained scenarios which represents a challenging obstacle for accurate face alignment. Different sources of occlusion are frequently seen covering the face such as: accessories (e.g. hats and glasses), beards, different hair styles, and self-occlusion due to extreme head poses. In spite of this, humans can quite accurately estimate a person's face shape while machine learning models often fail and produce unstable estimates. It is an important problem to address and has thus attracted the attention of the research community.

One of the first methods that explicitly handles occlusions was presented in [101] and was named Robust Cascaded Regression (RCPR). Burgoss-Artizzu et al. extended CPR [22] and ESR [14] methods and introduced a new, more challenging data set called Caltech Occluded Faces in the Wild (COFW) which has become a benchmark data set for face alignment with occlusions. It is publicly available and has annotations for occluded landmarks. At the beginning of the RCPR approach, the face image is divided into nine regions. At each stage multiple regressors (as in [14]) are trained, where each regressor is allowed to extract features from only one of the nine regions. Each such two-level boosted regressor learns to predict the occlusion of the corresponding region along with the landmark positions as the third dimension of the output vector. The predicted occlusion level from the previous stage is then used to assign the weights when combining current estimates of different regressors. This was the first method that also predicted occlusion with face alignment.

A different approach to occlusion handling was taken in [102]. The method was named Regional Predictive Power (RPP) and the main idea is to apply a graph-based segmentation of face images. The usefulness of each segment for face alignment task is determined by using the sieving regression forest votes [19]. A confidence value (if it is a face region) is obtained for each pixel, based on the center sieve from random forest votes. The RPP map is used to assign the weights to regressors from either of these methods: CPR [22], RCPR [101], ESR [14]. The weights were calculated by accumulating confidences from all the pixels used in those regressors.

Wu et al. use cascaded regression with explicit occlusion learning [103]. At each stage, landmark visibility probabilities are estimated first, then the landmark locations. The visibility probability updates at each stage are learned by extracting SIFT features around the current landmark positions and by concatenating with the shape features. The shape features are formed by calculating the difference between pairwise landmark locations. Additionally, an occlusion pattern loss term is added to the standard least squares objective function which penalizes improbable occlusion patterns. The loss function is constructed from auto-encoder network reconstruction errors. The landmark localization is learned in the same way and with the same features, however, the appearance part of the features are weighted by the visibility probabilities. Missing annotation is handled by adding binary weights to the weighted least squares problem.

While all of the above methods try to estimate occluded facial regions to avoid feature extraction from those regions, Zhang et al. propose a different approach in their work called Deep Regression Networks Coupled with De-corrupt AutoEncoders (DRDA) [104]. The aim is to reconstruct the occluded region using de-corrupt auto-encoders, again, in a cascaded regression framework. Deep regression and de-corrupt auto-encoder alternate each stage in order to benefit from each other. Estimated landmark positions are used to partition the face and the cropped image is fed into the auto-encoder to produce the un-occluded version which is then forwarded as input to the next deep regression stage. The method produces realistic images without occlusions which in turn improves alignment accuracy.

5.1. Summary

It is expected from face alignment methods to show a certain level of robustness to occlusions in order to be usable in real-life situations. The main approach in the literature is to estimate the level and region of occlusion in the image of the face in order to avoid extracting ambiguous features which cause unstable face shape predictions. However, with the recent introduction of generative models such as auto-encoders or Generative Adversarial Networks (GANs) [105], a new approach has emerged where the occluded region is reconstructed and used for accurate face alignment. An obvious drawback of this new direction is the increase in computational complexity. However, with optimized computing on GPUs, these methods demonstrate the potential to surpass human

Table 8
Occlusion modeling methods summary.

Methods	Architecture	Occlusion modeling	Year
RCPR [101]	Cascaded regression	Hand-crafted regions	2013
RPP [102]	Cascaded regression	Segmentation	2015
Wu et al. [103]	Cascaded regression	Landmark visibilities	2015
DRDA [104]	Deep cascaded regression	Decorrupt auto-encoder	2016

accuracy for occluded face alignment. A summary of the described methods is presented in Table 8.

6. Evaluation

This section provides a comparison of the described methods on widely used benchmark data sets. The results will be presented as reported in their respective papers. During the years of research, a number of different metrics have emerged to measure the alignment accuracy. The earliest and most frequently used is the normalized root mean squared error:

$$\frac{||S - S^*||_2}{D} \tag{10}$$

where D represents the normalization factor which varies between the following values in previous work:

- Inter-pupil distance (IPD) - this metric is the most common one, however, it can only be used on frontal or semi-frontal faces where pupils are visible [26].
- Inter-ocular distance (IOD) - this metric is used when there are no pupil annotations and represents the distance between outer corners of the eyes [106].
- Bounding box diagonal (BBD) - this metric is more suitable for profile faces where the first two metrics produce unreasonably small values [43].

The authors in [43] argue that the above average measure is not always informative enough since few outliers can affect the result significantly and propose Cumulative Error Distribution (CED) curves which provide a much more detailed source of information for analysis. Additional measurements from the curve are then derived:

- Area Under the Curve (AUC) - this metric represents the calculated area under the CED curve and is usually calculated up to a certain error threshold (e.g. 5% which is then marked with $AUC_{0.05}$).
- Failure Rate (FR) - this metric represents a percentage of samples with an error greater than a set threshold (e.g. 5% which is then marked with $FR_{0.05}$).

All of these metrics will be taken into account in addition to inference time of the proposed methods, if available, during the comparison. The following section will describe the data sets commonly used for both training and evaluation with additional emphasis on recent augmentation strategies to increase the training set.

6.1. Data sets

BioID data set was extensively used by older methods [107]. It consists of 1521 near-frontal face images captured in a laboratory environment and is, therefore, less challenging. The results on this data set are not reported because it has reached a saturation point. Additionally, methods usually only report cumulative error distribution curves so it is hard to compare between methods without reproducing the results.

300-W data set is currently adopted as the main benchmark data set for face alignment and is actually a compilation of different data sets (AFW, LFPW, HELEN, and XM2VTS) under consistent 68-point annotation [106]. It also contains a challenging set of 135 images called IBUG. A standard partitioning was set in [26] into a training set (the training set from LFPW and HELEN, whole AFW) with 3148 images and three testing sets: the test sets from LFPW and HELEN as the common subset, the whole IBUG as the challenging subset, and both common and challenging as the full test set with 689 images. The results on this data set are reported and analyzed in Section 6.2.

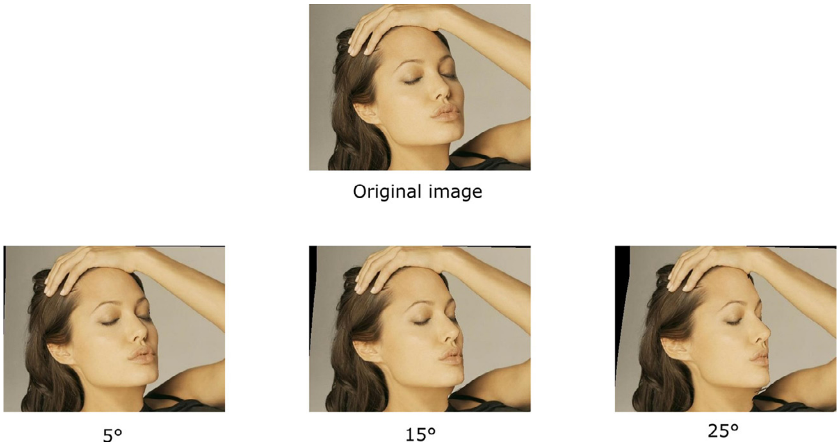
COFW data set was designed to present faces with occlusions due to pose, the use of accessories, and interaction with objects [101]. This data set includes 1007 faces, annotated with 29 landmarks. Methods that explicitly handle extreme poses and occlusions usually report results on this data set. The comparison is presented in Section 6.3.

HELEN data set contains 2000 training and 330 test images [108]. There are multiple different publicly available annotations for this data set. The results from this data set are not displayed because it is already included in 300-W data set.

LFPW data set originally contained 1100 training and 300 test images [109]. However, due to invalid URLs, a different number of images is available at different times so most methods use different images for training and testing and are, therefore, not fairly comparable. This is the main reason why results from this data set are not displayed. It is also included in 300-W data set.

Menpo data set was released as part of The Menpo Facial Landmark Localisation Challenge [43]. It contains 11k frontal/semi-frontal images and 3852 profile images. The aim was to increase the number of images and to add fully profile images compared to the 300-W data set. The frontal images are annotated with 68 landmarks in the standard IBUG marking setup and profile images have only visible landmarks annotated which adds up to 39 landmarks. The increased number of images facilitated deep learning research for face alignment. The results on this data set are presented in [43] and are, therefore, omitted from this work.

AFLW data set has a collection of 26 000 face images in unconstrained settings including a wide range of head poses which makes it especially suitable for large pose face alignment [110]. However, the data set is annotated with only 21 visible landmarks which limits its intended applicability. The **AFLW2000-3D** [81] data set was constructed from the first 2000 images of the AFLW data set. It contains a good distribution of “in-the-wild” head poses from frontal to full profile images with 3D annotations consistent with the 300-W data set. The annotations are generated using a 3DMM fitting procedure as described in Section 3, thus containing a full face shape with both visible and invisible landmarks. Another extension of the original AFLW data set was recently introduced as **AFLW-LPFA** with a balanced distribution of yaw angles (in total 1299 images) and additional 14 landmark annotations. Both extensions are currently used as benchmarks for large pose and 3D face alignment, therefore, we present the results and analysis on these data sets in Section 6.4.



Original image

5°

15°

25°

Fig. 11. Face profiling - from original image on the top to more extreme yaw rotations in the bottom row by applying additional rotation from left to the right [81].

6.1.1. Augmentation

As already emphasized multiple times, data set size is important in order to successfully train a face alignment model, especially using deep learning. However, it is extremely tedious to annotate a large number of landmarks accurately on an image. Therefore, multiple approaches were proposed during the years for data set augmentation. Standard random image manipulation techniques are already used in almost every previous work as part of data set pre-processing including: in-plane rotations, initial bounding box position and size perturbations, image mirroring etc. These are, however, useful up to a certain point and more complex techniques are required to artificially increase the variation and size of the data set.

Feng et al. in their Cascaded Collaborative Regression method (CCR) use synthetic images from 3D models in a standard cascaded regression approach with a dynamic annealing schedule for the mixing parameter that determines the ratio of synthesized and real images in the training set [111]. Greater emphasis is given to the synthesized images in earlier stages in order to make the prediction more robust with respect to various head poses.

Instead of generating synthetic images, Zhu et al. in [81] expanded the 300-W data set by warping the original images to increase the head pose variation and called it 300-W-LP. Unlike face frontalization [112] (but inspired by it), a 3DMM is fitted on original image annotations with the texture mapped to the model. The textured model is then rotated and rendered to generate extreme head pose samples with automatic annotations (see Fig. 11). The authors managed to increase the size of the data set by an order of magnitude to train their deep learning model.

Finally, automatic unification of differently annotated data sets is another logical option to increase the training data size. In [62], Bulat and Tzimiropoulos used the large 300-W-LP data set [81] with both 2D and 3D annotations to train a CNN that can translate from 2D to 3D. This model is then used to annotate all existing data sets to produce the largest 3D facial landmark data set (~ 230 000 images). Training a deep learning model on such a large data set resulted with state-of-the-art accuracy.

Table 9
Results reported on 300-W data set using IPD normalization. Methods marked with * use external data for training.

Method	Common	Challenging	Full
RCPR [101]	6.18	17.26	8.35
ESR [14]	5.28	17.00	7.58
SDM [24]	5.57	15.40	7.50
3DDFA [81]	6.15	10.59	7.01
CFAN [57]	5.50	-	-
RPP [102]	5.50	11.57	6.69
ERT [23]	-	-	6.40
LBF [26]	4.95	11.98	6.32
iMORF [95]	-	-	6.31
Jourabloo et al. [85]	5.43	9.88	6.30
PRF [27]	4.90	11.96	6.28
DRDA* [104]	-	10.79	-
DeFA* [86]	5.37	9.38	6.10
cGPRT [29]	-	-	5.71
CFSS [49]	4.73	9.98	5.76
R-DSSD [31]	4.16	9.20	5.59
ERT-PIS [32]	4.42	10.32	5.58
KRFWS [34]	4.62	9.48	5.57
MCNet [54]	-	8.87	-
TCDCN* [94]	4.80	8.60	5.54
DSRN* [52]	4.12	9.68	5.21
DAN [58]	4.42	7.57	5.03
DCR [36]	4.19	8.42	5.02
FEC-CNC [42]	4.20	7.90	4.90
DCFE [33]	3.83	7.54	4.55
DVLN* [51]	3.79	7.15	4.45
AWing [68]	3.77	6.52	4.31
LAB [65]	3.42	6.98	4.12
Wing* [59]	3.27	7.18	4.04
Yang et al.* [63]	-	7.0	-
Liu et al.* [66]	3.45	6.38	4.02

Table 10
Results reported on 300-W data set using IOD normalization. Methods marked with * use external data for training.

Method	Common	Challenging	Full	AUC_{thr}	FR_{thr}
LBF-NN [28]	4.08	10.30	5.26	-	-
P-DSC-CR* [37]	3.83	6.93	4.38	-	-
MDM [30]	-	-	4.05	$52.12_{0.08}$	$4.21_{0.08}$
KRFWS [34]	3.34	6.56	3.97	-	-
DAN [58]	3.19	5.24	3.59	$55.33_{0.08}$	$1.16_{0.08}$
LAB [65]	2.98	5.19	3.49	-	-
DeCaFA* [61]	2.93	5.26	3.39	$66.10_{0.1}$	$0.15_{0.1}$
HRNet [69]	2.87	5.15	3.32	-	-
DCFE [33]	2.76	5.22	3.24	$60.13_{0.08}$	$1.59_{0.08}$
Yang et al.* [63]	-	4.9	-	-	-
AWing [68]	2.72	4.52	3.07	-	-

6.2. Comparison on 300-W

As can be seen from Tables 9 and 10, the best results are achieved by heatmap regression methods using stacked hourglass architecture [63,65,66,68] and methods utilizing additional training data [51,59,61,63,66]. The two top-performing algorithms from Yang et al. [63] and Liu et al. [66] used two face detection algorithms that also output a subset of landmarks which were used to pre-align the faces for the stacked hourglass architecture training, making the alignment task less complex. The core contribution of the DVLN algorithm [51] is the use of additional training data (leveraging different mark-ups in data sets) from which every deep learning method should benefit. From the regression architecture stand-point, the stacked hourglass model demonstrates impressive results being featured in 4 out of 5 best performing methods. However, it suffers from high computational burden making real-time performance unfeasible even with a high-end GPU. Finally, there are two methods (Wing [59] and AWing [68]) among the top performers confirming the effectiveness of a customized loss compared to the standard L2 loss.

Using only 300-W data set, competitive results on both challenging and common subsets are achieved by lightweight DCFE [33] algorithm which utilizes global feature initialization through a CNN in combination with local cascaded regression (ERT [23]). The next three methods (FEC-CNC [42], DCR [36], and DAN [58]) use a combination of cascaded regression and deep learning to achieve good results but with a significant margin from the top performers. Methods DSRN [52] and TCDCN [94] both use single deep learning model and external data to achieve results lagging from the leading methods especially on the challenging subset. It seems that the coarse-to-fine approach from cascaded architectures is beneficial for deep learning models as well, on such a complex task.

Traditional cascaded regression methods (LBF [26], SDM [24], and ESR [14]), although revolutionary at their time and extremely efficient, can not compete with high capacity and flexibility of convolutional networks. The greatest problem for these methods and their derivatives are the challenging images with ground truth far away from the initial shape. Additionally, 3D alignment methods [81,85,86] also struggle on 2D benchmarks. Although highly robust as demonstrated on the challenging portion of the test set, disproportionately large errors are reported on the common portion. This is presumably due to the constraints imposed by the 3DMM as discussed in Section 3.

The Failure Rate is a bit more intuitive measure when evaluating face alignment accuracy. We can see that the two highly competitive methods DCFE [33] and DeCaFA [61] report $FR_{0.08} = 1.59\%$ and $FR_{0.1} = 0.15\%$, respectively. It has been known for some time that face alignment has been a solved problem in controlled environments, however, these results on benchmarks "in the wild" suggest that it is close to being solved in general as well.

6.3. Comparison on COFW

Table 11 presents the reported results on the COFW [101] data set. The novel HRNet [69] architecture employing high resolution representation achieves state-of-the-art results without utilizing additional training data. This demonstrates the effectiveness of high and low resolution representation fusion in the presence of occlusions. The effective use of larger training sets was the key driver of the second-best results achieved by LAB [65] algorithm. Boundaries are used as a middle-level representation of the ground-truth landmarks allowing the use of combined heterogeneous data sets to outperform other methods by a large margin. This is especially significant when training and testing on such a small data set as COFW (≈ 1000 faces) [101]. Third place is reserved for another heatmap regression method with stacked hourglass architecture but optimized with the custom AWing loss [68] demonstrating the effectiveness of the architecture without external data.

Surprisingly, competitive results are reported by Valle et al. with their DCFE [33] algorithm even though a simple approach of predicting landmark visibilities is used regarding occlusions. The probable reason for such high robustness towards occlusions are the final stages in their framework where the face shape is broken down into smaller semantic facial regions (e.g. mouth and eyes) which are regressed individually making them robust to features extracted from occluded regions. Another surprisingly good result is reported by Zhu et al. with their ERT-PIS [32] method by utilizing an improved initialization strategy for the ERT [23] method. This might be explained by the effectiveness of decision trees on small data sets such as COFW since less competitive results are reported on 300-W data set.

Table 11
Results reported on COFW data set. Methods marked with * use external data for training.

Method	Error (%)	AUC_{thr}	FR_{thr}
ESR [14]	11.22	-	$35.7_{0.1}$
SDM [24]	8.77	-	$24.32_{0.1}$
RCPR [101]	8.5	-	$20_{0.1}$
RPBM [48]	8.3	-	-
TCDCN* [94]	8.05	-	$15.31_{0.1}$
RPP [102]	7.52	-	$16.2_{0.1}$
R-CR-C [40]	7.3	-	$12.2_{0.1}$
CCR* [111]	7.03	-	$10.9_{0.1}$
DRDA* [104]	6.46	-	-
R-DSSD [31]	6.17	-	$8.23_{0.1}$
MCNet [54]	6.08	-	$2.96_{0.1}$
P-DSC-CR* [37]	6.06	-	$6.11_{0.1}$
Wu et al. [103]	5.93	-	-
Yang et al.* [63]	5.6	-	-
ERT-PIS [32]	5.54	-	-
DCFE [33]	5.27	$35.86_{0.08}$	$7.29_{0.08}$
AWing [68]	4.94	$39.11_{0.08}$	$5.52_{0.08}$
LAB* [65]	3.92	-	$0.39_{0.1}$
HRNet [69]	3.45	-	$0.19_{0.1}$

Following ERT-PIS [32] is another heatmap regression method from Yang et al. [63] using the large pre-trained stacked hourglass architecture in combination with supervised face transformation that involves two sophisticated face detectors. It can be argued that most methods can benefit from this technique and utilize recent advances in face detection algorithms.

Finally, the only other method that reports a mean error below 6% is the one from Wu et al. [103]. They explicitly handle occlusions and do not achieve such accuracy on other data sets which indicates poor generalization abilities. Nevertheless, the results are still impressive for a cascaded regression method without the use of convolutional features (SIFT is used). Other traditional cascaded regression methods report poor results with failure rates above 20%.

The COFW data set proved to be more difficult due to a higher percentage of heavily occluded faces. This can be seen by comparing failure rates for the same algorithm (DCFE) on the COFW ($FR_{0.08} = 7.29\%$) and 300W ($FR_{0.08} = 1.59\%$) data sets. If we take into account reported human performance on this data set with a mean error of 5.6%, we can see that several methods [33,65,68,69] already surpass it. However, humans do not make large errors with a reported failure rate of $FR_{0.1} = 0\%$ while HRNet reports $FR_{0.1} = 0.19\%$, meaning there is still room for improvements regarding face alignment robustness to occlusions.

6.4. Comparison on AFLW2000-3D and AFLW-LPFA

Both AFLW2000-3D and AFLW-LPFA data sets have problems with annotations, AFLW2000-3D due to the errors in the 3DMM fitting process while AFLW-LPFA has only visible annotations. However, these data sets are currently used as the main benchmarks for 3D and large pose face alignment due to the balanced distribution of head poses in unconstrained environment. An additional advantage of AFLW2000-3D is the partitioning of the test set according to the yaw rotation of the face providing additional insights.

Table 12
Results reported on AFLW2000-3D and AFLW-LPFA data sets. The protocol does not define the training set, however, most methods use 300-W-LP.

	AFLW2000-3D				AFLW-LPFA
Method	[0°, 30°]	[30°, 60°]	[60°, 90°]	Mean	Mean
PIFA [83]	-	-	-	-	8.04
RCPR [101]	4.26	5.96	13.18	7.80	6.26
Yu et al. [91]	3.62	6.06	9.56	-	-
3DDFA [81]	3.78	4.54	7.93	5.42	-
3DDFA + SDM [81]	3.43	4.24	7.17	4.94	-
PAWF [84]	-	-	-	-	4.72
Jourabloo et al. [85]	-	-	-	-	4.45
DeFA [86]	-	-	-	4.50	3.86
3DSTN [88]	3.15	4.33	5.98	4.49	-
PRNet [92]	2.75	3.51	4.61	3.62	2.93
CMHM [82]	2.36	2.80	4.08	3.08	-

The results on both data sets are presented in Table 12 using the standard error measurement (Eq. 10) normalized by the bounding box diagonal. It is evident from the results on AFLW2000-3D that all methods suffer from a drop in accuracy when moving from frontal to profile head pose which is understandable since it is difficult to both annotate and estimate self-occluded landmarks.

The best result on AFLW2000-3D data set is reported by Deng et al. with their CMHM method [82]. The critical difference is the additional use of joint multi-view 2D supervision. As already mentioned, the 3D annotations are constrained with the flexibility of the 3DMM and can not represent the real “ground truth” as do the manually annotated 2D landmarks. The other two top contenders, PRNet [92] and 3DSTN [88], also implement techniques to mitigate explicit constraints of the 3DMM even though they still exist implicitly through the data set (300-W-LP).

In general, the 3D face alignment has also reached high levels of precision across the full range of head poses by utilizing a large body of previous work for 2D face alignment. If we factor in the different normalization factors, the state-of-the-art accuracy on AFLW2000-3D and AFLW-LPFA are quite similar to the best results on 300-W and COFW data sets.

6.5. Comparison of execution times

Tables 13 and 14 present execution times in milliseconds on CPUs and GPUs respectively for methods that report them in their publications. It is immediately visible that all fast methods with frame rates above 60 FPS on a single CPU core utilize simple and fast features such as Pixel Difference Features (PDF) or SIFT. The fastest method reporting 1 ms execution time is the ERT [23] method which uses PDF and fast decision trees. Other similar methods such as LBF [26] and LBF-NN [28] also report similar times with additional lighter and hyper-fast versions operating at above 3000 FPS. Unfortunately, these methods demonstrate significantly lower robustness and accuracy as can be seen in Sections 6.2 and 6.3.

Any method utilizing large CNNs without optimizations can achieve real-time performance only by employing high-end GPU hardware. Notable examples are DAN [58] and FAN [62] algorithms with execution times of 22.22 ms and 34 ms, respectively. The recent heatmap regression methods are incapable of real-time performance even on a GPU and, thus, mostly do not report execution time with an exception of LAB [65] (60 ms). However, recent work in [52] (DSRN) demonstrates impressive execution time of 2 ms on a GPU by utilizing more efficient convolutional layers [53]. Another example of a CNN architecture optimization, this time by reducing the complexity of the problem and thus required CNN complexity, is demonstrated in DVLN [51] with reported execution time of 15.15 ms on a CPU.

The method with both competitive accuracy and real-time performance is DCFE [33] reporting execution time of 31.25 ms on a CPU. It utilizes a heavy CNN architecture for global initialization which can probably be optimized by employing some of the ideas from DSRN [52] and DVLN [51] to make it even faster and more suitable for mobile and embedded platforms.

Table 13
Reported execution times in milliseconds on a CPU.

Method	Device	Exec. time
Wu et al. [103]	-	500
iMORF [95]	Core i7 @ 3.6 GHz	350
RCPR [101]	CPU @ 3.47 GHz	333.33
RPP [102]	CPU @ 3.3 GHz	250
P-DSC-CR [37]	-	100
FEC-CNC [42]	-	100
R-CR-C [40]	CPU @ 3.0 GHz	45.45
CFSS pract[49].	Core i5-4590	40
PRF [27]	Core i7-2600	33.33
DCFE [33]	Xeon E5-1650 @ 3.5 GHz	31.25
R-DSSD [31]	Core i5-6500 @ 3.2 GHz	25
TCDCN [94]	Core i5	18
MCNet [54]	Core i5-6200U	18
DVLN [51]	Core i5-4300u	15.15
CCR [111]	CPU @ 3.5 GHz	14.49
SDM [24]	Core i7-2600	14.3
cGPRT [29]	Core i5-3570 @ 3.4 GHz	10.75
ESR [14]	Core i7-2600	8.34
ERT-PIS [32]	Core i5-3470 @ 3.2 GHz	4.48
LBF [26]	Core i7-2600	3.12
LBF-NN [28]	Core i7-2600 @ 3.4 GHz	1.43
ERT [23]	-	1

Table 14
Reported execution times in milliseconds on a GPU.

Method	Device	Exec. time
Jourabloo et al. [85]	GTX Titan X	232.5.1
3DDFA [81]	GTX Titan Black	75.72
LAB [65]	GTX Titan X	60
FAN [62]	GTX Titan X	34
DeCaFA [61]	GTX 1060	31.25
DAN [58]	GTX 1070	22.22
3DSTN [88]	GTX Titan X	19
Yu et al. [91]	GTX Titan X	19
PRNet [92]	GTX 1080	9.8
Wing [59]	GTX Titan X	5.88
DSRN [52]	GTX 1080Ti	2

6.6. Discussion

Taking into account accuracies on 2D benchmark data sets (without external training data) and real-time performance on a CPU, the overall state-of-the-art method is DCFE [33] from Valle et al. It demonstrates both high robustness and fine accuracy, terms that are usually inversely proportional. Key features of the method will be discussed in the following paragraphs.

Deep Global Initialization

As can be seen from other methods [42,58,94], CNNs are well suited for coarse face alignment (visible from the errors on the challenging 300-W subset). This is due to a couple of reasons. Firstly, CNNs use global features taking into account the whole face holistically and the context as well. This makes it easier to infer the global orientation of the face and head. Secondly, CNNs posses high capacity and flexibility to absorb extreme variations in appearance due to different head poses and backgrounds.

Shape-indexed local features

It can be seen from the large number of methods adopting cascaded regression in Section 2.2 that gradual alignment helps to improve both accuracy and robustness. However, local shape-indexed features in later stages of the cascade seem to be important as well by providing the algorithm an attention mechanism necessary for fine-grained alignment. For instance, Kowalski et al. used global CNNs in all stages of the cascade, however, in order to achieve competitive results a heatmap constructed from current landmark positions needed to be passed to the next stage of the cascade. It helped later stages to “focus” on relevant regions of the face. Even so, the method achieved significantly worse results on the common 300-W subset than methods utilizing local shape-indexed features (DCFE [33], FEC-CNC [42], DCR [36]) while maintaining similar results on the challenging subset.

Part-based fine-tuning

This is the key technique that sets DCFE apart from other cascaded regression methods. The last stages in the DCFE cascade do not regress a single monolithic face shape. It is broken up into semantic facial parts consisting of landmarks relevant for that region (e.g. eyes, mouth, nose etc.). Even though early face alignment attempts used a similar approach [16], the important difference here is that it is used at the end of the cascade with landmarks already close to the ground truth positions. This enables the method to accurately align asymmetrical facial expressions not seen in the training set due to large number of possible part combinations.

In addition to the described techniques, it is evident from the comparisons that an increase in training data size improves accuracy, especially for methods utilizing deep learning [51,63]. It is understandable since the size of the data sets are still quite low (300-W training set has 3148 images) due to the high complexity of the annotation process. Techniques described in Section 6.1.1 are thus interesting to explore and use in combination with highly efficient algorithms [33,42,58]. Similar results can be observed on the most recent face alignment Menpo challenge with more details in [43].

In order to achieve large pose face alignment, 3D information needs to be introduced in the design of a robust face alignment model. As already discussed, 3D data sets are still inadequate to be used alone. A simple solution is to combine manually annotated visible 2D landmarks and generate annotations for the self-occluded landmarks utilizing the 3DMM fitting procedure. In such manner, both high precision in frontal pose and high robustness to extreme rotations can be achieved.

7. Conclusion

Face alignment remains one of the most important and basic problems in face analysis with faces one of the most interesting objects to observe in computer vision. Recent advances in face alignment have been presented, analyzed, and compared in this work. The most successful methods utilize a hybrid approach with global CNNs and local shape-indexed features organized in a cascaded regression framework. This architecture enables high levels of robustness and precision in unconstrained environments.

The question arises: is the problem solved then? Indeed, these algorithms surpass human-level precision even on data sets “in the wild”. However, failure rates still indicate insufficient robustness. Human annotators, although not as precise, still do not fail regardless of the pose, occlusion, or lighting conditions. This is very important in industries such as automotive where such failures can have dire consequences (e.g. driver monitoring). In order to surpass human-level precision and robustness, 3D data sets will need to be improved and included in the face alignment pipeline.

Additionally, computational efficiency is becoming increasingly important since face tracking in videos is a dynamic task with many applications requiring low latency. In order to achieve real-time performance on devices with limited resources, which are becoming prevalent in the IoT world, the algorithms will need to be optimized. Fortunately, the research community has recognized this trend with increasing attention for deep learning optimization topics.

Generative models have recently exploded in the research community with their ability to generate artificial photo-realistic images. One of the useful applications is the automatic creation of large-scale data sets which would be especially beneficial for face alignment. Another promising use of generative models already being researched is for face alignment under heavy occlusions with the potential to “see” the parts of face behind the obstacle.

Since face alignment is closely related to face detection, expressions, age, gender, and other face analysis tasks, it makes sense to unify the predictive models under a holistic approach. Past research confirms the merits of multi-task learning for related problems. The largest obstacle, however, is the unification of divergent data sets with different annotations. As already mentioned, generative models could be used to produce a well balanced large-scale data set to train a holistic face analysis model.

To conclude, even though face alignment seems to be solved in most scenarios, there are still interesting research directions that could produce accuracy and stability beyond human level. This survey can serve as a starting point for further research in this interesting field.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Acknowledgments

This research is funded by Visage Technologies AB.

References

[1] E. Hjelmås, B.K. Low, Face detection: a survey, Comput. Vision Image Understanding 83 (3) (2001) 236–274.

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

[3] V. Bettadapura, Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722.

[4] A.K. Jain, S.Z. Li, Handbook of face recognition, 1, Springer, 2005.

[5] A.F. Abate, M. Nappi, D. Riccio, G. Sabatino, 2D and 3d face recognition: a survey, Pattern Recognit. Lett. 28 (14) (2007) 1885–1906.

[6] J. Shen, S. Zafeiriou, G.G. Chrysos, J. Kossaifi, G. Tzimiropoulos, M. Pantic, The first facial landmark tracking in-the-wild challenge: Benchmark and results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 50–58.

[7] I. Gogić, M. Manhart, I.S. Pandžić, J. Ahlberg, Fast facial expression recognition using local binary features and shallow neural networks, Vis. Comput. (2018) 1–16.

[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.

[9] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, S. Zafeiriou, Large scale 3d morphable models, Int. J Comput. Vis. 126 (2–4) (2018) 233–254.

[10] X. Jin, X. Tan, Face alignment in-the-wild: a survey, Comput. Vision Image Understanding 162 (2017) 1–22.

[11] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, Proceedings of the European Conference on Computer Vision 2 (1998) 484–498, doi:10.1109/34.927467.

[12] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164, doi:10.1023/B:VISI.0000029666.37597.d3.

[13] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models-Their training and application, Comput. Vision Image Understanding 61 (1) (1995) 38–59, doi:10.1006/cviu.1995.1004.

[14] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2012).

[15] M. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (2010) 2729–2736, doi:10.1109/CVPR.2010.5539996.

[16] V. Kazemi, J. Sullivan, Face alignment with part-Based modeling, British Machine Vision Conference (2011) 27.1–27.10, doi:10.5244/C.25.27.

[17] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, 2007, pp. 401–408.

[18] H. Yang, I. Patras, Face parts localization using structured-output regression forests., in: ACCV (2), 2012, pp. 667–679.

[19] H. Yang, I. Patras, Sieving regression forest votes for facial feature detection in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1936–1943.

[20] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection., IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163, doi:10.1109/TPAMI.2012.205.

[21] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7) (2002) 971–987.

[22] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010) 1078–1085, doi:10.1109/CVPR.2010.5540094.

[23] V. Kazemi, J. Sullivan, One Millisecond Face Alignment with an Ensemble of Regression Trees, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1867–1874, doi:10.1109/CVPR.2014.241.

[24] X. Xiong, F. De La Torre, Supervised descent method and its applications to face alignment, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2013) 532–539, doi:10.1109/CVPR.2013.75.

[25] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[26] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.

[27] C. Luo, Z. Wang, S. Wang, J. Zhang, J. Yu, Locating facial landmarks using probabilistic random forest, Signal Processing Letters, IEEE 22 (12) (2015) 2324–2328.

[28] N. Markuš, I. Gogić, I.S. Pandžić, J. Ahlberg, Memory-efficient global refinement of decision-tree ensembles and its application to face alignment, in: British Machine Vision Conference BMVC, 2018.

[29] D. Lee, H. Park, C.D. Yoo, Face alignment using cascade Gaussian process regression trees, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 4204–4212, doi:10.1109/CVPR.2015.7299048.

[30] G. Trigeorgis, P. Snape, M.A. Nicolaou, E. Antonakos, S. Zafeiriou, Mnemonic descent method: A recurrent process applied for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4177–4187.

[31] H. Liu, J. Lu, J. Feng, J. Zhou, Learning deep sharable and structural detectors for face alignment, IEEE Trans. Image Process. 26 (4) (2017) 1666–1678.

[32] H. Zhu, B. Sheng, Z. Shao, Y. Hao, X. Hou, L. Ma, Better initialization for regression-based face alignment, Computers & Graphics 70 (2018) 261–269.

[33] R. Valle, J.M. Buenaposada, A. Valdes, L. Baumela, A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 585–601.

[34] M. Kowalski, J. Naruniec, Face alignment using k-cluster regression forests with weighted splitting, IEEE Signal Process. Lett. 23 (11) (2016) 1567–1571.

[35] K. Hara, R. Chellappa, Growing regression forests by classification: Applications to object pose estimation, in: European conference on computer vision, Springer, 2014, pp. 552–567.

[36] H. Lai, S. Xiao, Z. Cui, Y. Pan, C. Xu, S. Yan, Deep cascaded regression for face alignment. arXiv preprint arXiv:1510.09083.

[37] Q. Liu, J. Deng, D. Tao, Dual sparse constrained cascade regression for robust face alignment., IEEE Trans Image Process 25 (2) (2016) 700–712, doi:10.1109/TIP.2015.2502485.

[38] M. Aharon, M. Elad, A. Bruckstein, K-Svd: an algorithm for designing overcomplete dictionaries for sparse representation, Signal Processing, IEEE Transactions on 54 (11) (2006) 4311–4322.

[39] J. Yan, Z. Lei, D. Yi, S. Li, Learn to combine multiple hypotheses for accurate face alignment, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 392–396.

[40] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, X.-J. Wu, Random cascaded-regression copse for robust facial landmark detection, IEEE Signal Process. Lett. 1 (22) (2015) 76–80.

[41] Z. He, J. Zhang, M. Kan, S. Shan, X. Chen, Robust fec-cnn: A high accuracy facial landmark detection system, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 98–104.

[42] Z. He, M. Kan, J. Zhang, X. Chen, S. Shan, A fully end-to-end cascaded cnn for facial landmark detection, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 200–207.

[43] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen, The menpo facial landmark localisation challenge: A step towards the solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 170–179.

[44] X. Xiong, F. De la Torre, Global supervised descent method, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 2664–2673, doi:10.1109/CVPR.2015.7298882.

[45] S. Zhu, C. Li, C.-C. Loy, X. Tang, Unconstrained face alignment via cascaded compositional learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3409–3417.

[46] Y. Dong, Y. Wang, J. Yue, Z. Hu, Robust facial landmark localization using multi partial features, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 98–102, doi:10.1109/ICIP.2015.7350767.

[47] S. Zhu, C. Li, C.C. Loy, X. Tang, Towards arbitrary-view face alignment by recommendation trees. arXiv preprint arXiv:1511.06627.

[48] K. Rampal, K. Sakurai, H. Imaoka, Occlusion handling in feature point tracking using ranked parts based models, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 740–744, doi:10.1109/ICIP.2015.7350897.

[49] Shizhan Zhu, Cheng Li, C.C. Loy, X. Tang, Face alignment by coarse-to-fine shape searching, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 4998–5006, doi:10.1109/CVPR.2015.7299134.

[50] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: European conference on computer vision, Springer, 2010, pp. 778–792.

[51] W. Wu, S. Yang, Leveraging intra and inter-dataset variations for robust face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 150–159.

[52] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, H. Huang, Direct shape regression networks for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5040–5049.

[53] S. Zhai, Y. Cheng, Z.M. Zhang, W. Lu, Doubly convolutional neural networks, in: Advances in neural information processing systems, 2016, pp. 1082–1090.

[54] Z. Shao, H. Zhu, Y. Hao, M. Wang, L. Ma, Learning a multi-center convolutional network for unconstrained face alignment, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 109–114.

[55] Y. Sun, X. Wang, X. Tang, Deep Convolutional Network Cascade for Facial Point Detection, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 3476–3483, doi:10.1109/CVPR.2013.446.

[56] , Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade, in: 2013 IEEE International Conference on Computer Vision Workshops, IEEE, 2013, pp. 386–391, doi:10.1109/ICCVW.2013.58.

[57] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine Auto-encoder Networks (Cfan) for Real-time Face Alignment, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 1–16.

[58] M. Kowalski, J. Naruniec, T. Trzcinski, Deep alignment network: A convolutional neural network for robust face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 88–97.

[59] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, X.-J. Wu, Wing loss for robust facial landmark localisation with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2235–2245.

[60] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in neural information processing systems, 2015, pp. 2017–2025.

[61] A. Dapogny, K. Bailly, M. Cord, Decafa: Deep convolutional cascade for face alignment in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6893–6901.

[62] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[63] J. Yang, Q. Liu, K. Zhang, Stacked hourglass network for robust facial landmark localisation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 79–87.

[64] D. Chen, G. Hua, F. Wen, J. Sun, Supervised transformer network for efficient face detection, in: European Conference on Computer Vision, Springer, 2016, pp. 122–138.

[65] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, Q. Zhou, Look at boundary: A boundary-aware face alignment algorithm, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2129–2138.
[66] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N.M. Robertson, J. Wang, Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3467–3476.
[67] L. Chen, H. Su, Q. Ji, Face alignment with kernel density deep neural network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6992–7002.
[68] X. Wang, L. Bo, L. Fuxin, Adaptive wing loss for robust face alignment via heatmap regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6971–6981.
[69] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[70] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, Facewarehouse: a 3d facial expression database for visual computing, IEEE Trans. Vis. Comput. Graph. 20 (3) (2013) 413–425.
[71] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3d facial expression database for facial behavior research, in: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, 2006, pp. 211–216.
[72] L. Yin, X.C.Y. Sun, T. Worm, M. Reale, A high-resolution 3d dynamic facial expression database, 2008, in: IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, volume 126, p. 6.
[73] P.J. Besl, N.D. McKay, Method for registration of 3-d shapes, in: Sensor fusion IV: control paradigms and data structures, 1611, International Society for Optics and Photonics, 1992, pp. 586–606.
[74] B. Amberg, S. Romdhani, T. Vetter, Optimal step nonrigid icp algorithms for surface registration, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[75] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, T. Vetter, Morphable face models-an open framework, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 75–82.
[76] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, S. Zafeiriou, 3d face morphable models" in-the-wild", in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5464–5473.
[77] X. Zhu, Z. Lei, J. Yan, D. Yi, S.Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 787–796.
[78] S. Tulyakov, N. Sebe, Regressing a 3d face shape from a single image, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3748–3755.
[79] Y. Wu, Q. Ji, Shape augmented regression method for face alignment, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 26–32.
[80] R. Zhao, Y. Wang, C.F. Benitez-Quiroz, Y. Liu, A.M. Martinez, Fast and precise face alignment and 3d shape reconstruction from a single 2d image, in: European Conference on Computer Vision, Springer, 2016, pp. 590–603.
[81] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: A 3d solution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 146–155.
[82] J. Deng, Y. Zhou, S. Cheng, S. Zaferiou, Cascade multi-view hourglass model for robust 3d face alignment, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 399–403.
[83] A. Jourabloo, X. Liu, Pose-invariant 3d face alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3694–3702.
[84] A. Jourabloo, X. Liu, Pose-invariant face alignment via cnn-based dense 3d model fitting, Int. J. Comput. Vis. 124 (2) (2017) 187–203.
[85] A. Jourabloo, M. Ye, X. Liu, L. Ren, Pose-invariant face alignment with a single cnn, in: Proceedings of the IEEE International Conference on computer vision, 2017, pp. 3200–3209.
[86] Y. Liu, A. Jourabloo, W. Ren, X. Liu, Dense face alignment, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1619–1628.
[87] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.
[88] C. Bhagavatula, C. Zhu, K. Luu, M. Savvides, Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3980–3989.
[89] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell. 11 (6) (1989) 567–585.
[90] A.S. Jackson, A. Bulat, V. Argyriou, G. Tzimiropoulos, Large pose 3d face reconstruction from a single image via direct volumetric cnn regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1031–1039.
[91] R. Yu, S. Saito, H. Li, D. Ceylan, H. Li, Learning dense facial correspondences in unconstrained images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4723–4732.
[92] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3d face reconstruction and dense alignment with position map regression network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 534–551.
[93] T. Evgeniou, M. Pontil, Regularized multi–task learning, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 109–117.
[94] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial Landmark Detection by Deep Multi–task Learning, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 94–108.
[95] X. Zhao, T.-K. Kim, W. Luo, Unified Face Analysis by Iterative Multi-output Random Forests, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1765–1772, doi:10.1109/CVPR.2014.228.
[96] M. Dantone, J. Gall, G. Fanelli, L. Van Gool, Real-time facial feature detection using conditional regression forests, 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012) 2578–2585, doi:10.1109/CVPR.2012.6247976.
[97] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, in: European Conference on Computer Vision, Springer, 2014, pp. 109–122.
[98] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[99] J. Deng, G. Trigeorgis, Y. Zhou, S. Zafeiriou, Joint multi-view face alignment in the wild, IEEE Trans. Image Process.. 28 (7) (2019) 3636–3648.
[100] Y. Zhao, F. Tang, W. Dong, F. Huang, X. Zhang, Joint face alignment and segmentation via deep multi-task learning, Multimed. Tools Appl. 78 (10) (2019) 13131–13148.
[101] X.P. Burgos-Artizzu, P. Perona, P. Dollar, Robust face landmark estimation under occlusion, Proceedings of the IEEE International Conference on Computer Vision (2013) 1513–1520, doi:10.1109/ICCV.2013.191.
[102] Heng Yang, Xuming He, Xuhui Jia, I. Patras, Robust face alignment under occlusion via regional predictive power estimation., IEEE Trans Image Process 24 (8) (2015) 2393–2403, doi:10.1109/TIP.2015.2421438.
[103] Y. Wu, Q. Ji, Robust Facial Landmark Detection Under Significant Head Poses and Occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3658–3666.
[104] J. Zhang, M. Kan, S. Shan, X. Chen, Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3428–3437.
[105] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[106] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.
[107] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz, Robust face detection using the hausdorff distance, in: Audio-and video-based biometric person authentication, Springer, 2001, pp. 90–95.
[108] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive Facial Feature Localization, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 679–692.
[109] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (12) (2013) 2930–2940.
[110] M. Koestinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, 2011, pp. 2144–2151.
[111] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, X.-J. Wu, Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting., IEEE Trans Image Process 24 (11) (2015) 3425–3440, doi:10.1109/TIP.2015.2446944.
[112] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4295–4304.