Johnston and Chazal *EURASIP Journal on Image and Video Processing* (2018) 2018:86
https://doi.org/10.1186/s13640-018-0324-4

EURASIP Journal on Image and Video Processing

REVIEW Open Access

A review of image-based automatic facial landmark identification techniques

Benjamin Johnston[1,2*] and Philip de Chazal[1]

Abstract

The accurate identification of landmarks within facial images is an important step in the completion of a number of higher-order computer vision tasks such as facial recognition and facial expression analysis. While being an intuitive and simple task for human vision, it has taken decades of research, an increase in the availability of quality data sets, and a dramatic improvement in computational processing power to achieve near-human accuracy in landmark localisation. The intent of this paper is to provide a review of the current facial landmarking literature, outlining the significant progress that has been made in the field from classical generative methods to more modern techniques such as sophisticated deep neural network architectures. This review considers a generalised facial landmarking problem and provides experimental examples for each stage in the process, reporting repeatable benchmarks across a number of publicly available datasets and linking the results of these examples to the recently reported performance in the literature.

Keywords: Face, Landmarking, Registration, Image, Vision, Machine learning, Deep learning, Artificial neural networks, Review, Survey

1 Introduction

The accurate identification of specific facial features and landmarks is a foundational process by which a number of more complicated image analysis problems are solved. Tasks such as facial identification, expression analysis, age estimation, and gender classification are often built upon a facial landmarking component in their methods [1, 2]. The use of image-based automated facial landmarking has been extended outside of the domain of image research and into other applications, including some within the medical field. Conditions such as facial palsy, facial paralysis, and even sleep apnoea are either characterised by or associated with unique facial structures that enables the use of facial landmarking as a useful research or even screening tool. Very recently, Guarin et al. [3] described an automated facial landmarking tool which is used in the characterisation of facial displacements in sufferers of facial palsy; while the work by Anping et al. [4] uses landmarks as predicted by Active Shape Models to assess facial nerve paralysis. The use of facial landmarking methodologies has been recently examined as a means of screening for sleep apnoea by Tabatabaei Balaei et al. [5], looking at the association between the underlying structure of predicted facial landmarks and the likelihood of suffering from obstructive sleep apnoea. The authors of this review have also investigated the use of facial landmarks in sizing sleep apnoea masks [6], a critical device in sleep apnoea treatment. Given the wide variety of applications in which facial landmarking is applied, and in the case of medical applications the critical nature of the tasks it is vital that the systems be capable of accurately identifying the landmarks of interest.

While the process of identifying features such as the corner of an eye on a face is a natural and instinctive task for human vision; it has proven somewhat more challenging for computer vision, which has not benefited from millenia of evolution. Despite the overall similarity in the general content of facial images, common differences such as variation in pose, lighting, facial expression, and variations in the facial features themselves can be problematic for many computer vision systems leading to significant errors in landmarking accuracy.

*Correspondence: ben.johnston@sydney.edu.au
[1]Sleep Research Group, Charles Perkins Centre, School of Electrical and Information Engineering, 2006 University of Sydney, Sydney, NSW, Australia
[2]ResMed Ltd, 1 Elizabeth Macarthur Dr, 2153 Bella Vista, NSW, Australia

The intent of this paper is to review the current state of automated image-based facial landmarking processes and provide a comparison of the performance achieved by some of these methods. This paper aims to build upon the comprehensive review completed by Çeliktutan et al. [7]. Since the publication of this article in 2013 increases in computing power through the reduction in the cost of GPUs, in addition to an increase in the availability of large datasets has enabled the development of highly accurate, though computationally expensive methods. While not covered within this review, readers may also be interested in the automated detection of facial landmarks in three-dimensional models. With the improved availability and reduced cost of three dimensional scanners such as the Kinect or even those found in late model smartphones, 3D facial models are more readily available for analysis. While in some respects, automated 3D facial landmarking has evolved from its 2D counterpart, significant differences exist in the current focus of research. Three-dimensional landmarking currently uses a series of powerful geometric descriptors to both interpret and summarises the complex information encoded within the 3D model. Marcolin and Vezzetti describe 105 novel descriptors [8] mapped on 217 facial depth maps to generate a set of landmarks within a number of different facial expressions. Similar descriptors used by Vezzetti et al. in 2017 [9] achieved a mean localization error of 4.75 mm in facial scans containing differing emotions as well as the presence of occlusions, such as fingers or hands covering the face.

The structure of this review will differ somewhat to previous image processing surveys. This discussion will occur within the framework of a generic process for constructing an automated facial landmarking model; reviewing the state of the art at each stage of model construction. The details of this structure will be outlined in more detail in the next section; however, it is intended that reviewing the literature in the framework of a generic process will provide the reader with more clarity of the progress that has been made within the field. To further support the review, stages 2, 3, and 4 also contain experimental components with the intent of improving reader understanding of the corresponding stage. For a high performing, automated landmarking algorithm, it is important that each stage in constructing an automated methodology be carefully considered and appropriate design choices made. By discussing the state of the art within the context of these stages, the authors hope the reader is able to more completely understand the current state of the art.

2 Review

2.1 Generic model construction

While there are many differences between the various applications and methods of automated facial landmarking, it is possible to describe a generic process by which almost all models are created. This process which is not necessarily unique to facial landmarking provides an effective means for comparing different methodologies. We will define the generic model construction process as comprising five stages:

1. Definition of the objective: what is the exact nature of the problem to be solved?,
2. Selection of an appropriate dataset for solving the defined problem: what information is required to meet the objective?,
3. Extraction of regions of interest from the dataset: what features from the selected dataset will best meet the defined objective?,
4. Definition of model architecture: which model will give the best performance? and
5. Model training and evaluation: what is the best training methodology given all of the above stage?

Each of the following sections will discuss in detail each stage of the generic model construction process.

2.2 Stage 1: objective definition

The objective definition is arguably the most important step in the model construction process. A clear, concise, and correct definition of what is to be achieved is crucial as it forms the basis of all other steps, design decisions, and is often a platform for solving more complex problems. Wu et al. in 2012 [1] used facial landmarks to assist in age estimation and face verification; Devries et al. [2] used landmarking in facial expression recognition while Tabatabaei Balaei et al. [5] investigated the use of facial landmarks as a means of determining the likelihood that an individual sufferer from obstructive sleep apnoea (OSA).

This review paper will not consider any higher level applications and will define the objective definition for the generic model construction process as

> Aiming to construct an automatic facial landmarking system with performance comparable to that of an expert human annotator

2.2.1 Measuring performance

Performance metrics will vary depending upon the objective definition; Devries et al. [2] used expression classification accuracy while Tabatabaei Balaei et al. [5] measured performance based on rates of correct OSA diagnosis. While these measures determine overall system performance, measuring facial landmarking accuracy is required to ensure landmark predictions are acceptable. This is completed by comparing the predictions made by the system to a set of 'ground truth' landmarks which have been manually annotated by one or more human experts (see Figs. 1 and 2).

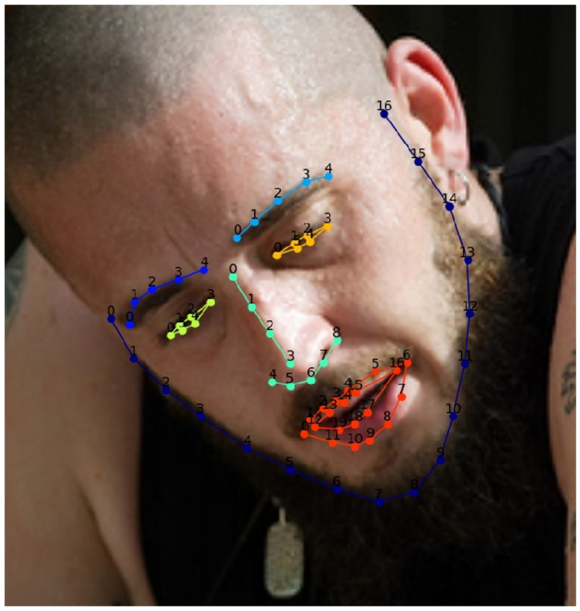

Fig. 1 Example annotated image. Example training image provided for the second facial landmark localisation competition held in conjunction with International Conference on Computer Vision and Pattern Recognition (CVPR) 2017 [57]

The simplest comparison is a root mean squared error (RMSE) assessment; where the average distance between each of the N predicted landmarks (x_i^p, y_i^p) and the corresponding 'ground truth' (x_i^t, y_i^t) is calculated on a per landmark basis. Landmarks that are poorly predicted will be positioned far their corresponding ground truth locations and thus contribute to increasing the RMSE value. Often the root mean squared error is normalised by the distance between two specific 'ground truth' points (NMRSE) such as the left (x_{le}^t, y_{le}^t) and the right (x_{re}^t, y_{re}^t) outer corners of the eyes d_{norm} (see Eq. 2) [10] to allow a fair comparison between faces of different sizes

$$\text{RMSE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(x_i^p - x_i^t\right)^2 + \left(y_i^p - y_i^t\right)^2}, \tag{1}$$

$$\text{NRMSE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{\left(x_i^p - x_i^t\right)^2 + \left(y_i^p - y_i^t\right)^2}}{d_{\text{norm}}}, \tag{2}$$

$$\text{d}_{\text{norm}} = \sqrt{\left(x_{le}^t - x_{re}^t\right)^2 + \left(y_{le}^t - y_{re}^t\right)^2} \tag{3}$$

When comparing the performance of different landmarking algorithms against the same dataset, the average RMSE or NRMSE value over the number of samples in the dataset (K) may simply be reported. A more detailed summary of the model performance can be provided using the cumulative error distribution (CED), which plots the cumulative NRMSE against the proportion of images with an NRMSE of less than or equal to a particular value (e.g. Figs. 11, 12, 13, 14, 15, 16, 17, and 18).

A less frequently reported metric is the landmark detection rate, i.e. the proportion of the N landmarks from the K images, correctly identified by the system. A landmark is correctly identified if its position is less than a defined Euclidean distance from the 'ground truth'. Similarly to the mean squared error calculations, landmark detection rate can also occur on a per-image, per-landmark, and overall average basis.

> Throughout this review paper we will use normalised root mean squared error (NRMSE), providing point-to-point CED plots to compare the performance of different landmarking methodologies.

2.3 Stage 2: dataset selection

A correct and appropriate selection of a dataset is crucial for the development of any predictive algorithm. The selected dataset must contain features with sufficient predictive power that the training process can 'learn' the relationships within the data. For many problems, such as the prediction of obstructive sleep apnoea [5] a custom dataset is required which itself may be subject to iterations of improvement to ensure the most appropriate features are being used.

For many facial analysis problems, there exists large, publicly available databases with rich feature sets (see Table 1). The datasets can be divided into two categories: those produced within a controlled environment and those produced in an uncontrolled environment. The development of social networks such as Facebook and Flickr, image search engines such as Google Images and the ability to obtain images at mass from these sites has enabled the construction of large datasets of facial images in various, uncontrolled situations. These 'in-the-wild' datasets have proven vital for facial landmarking/analysis problems where it is important to achieve high levels of accuracy without the burden of maintaining a controlled environment.

While many datasets are available for use, it can be seen in Table 1 that there is little consistency amongst the different sets. They have different numbers of samples, image resolutions, and configurations of ground truth landmarks. Some sets have multiple subjects in some images, while others have multiple images for some subjects. Many in-the-wild datasets are built from web links which may not still be valid. While such data variety is useful, it can lead to difficulties in comparing the performance of facial landmarking algorithms. Unlike image classification problems which often state performance using standard reference datasets such as MNIST and CIFAR, facial landmarking literature has not benefited from a common means of comparison.

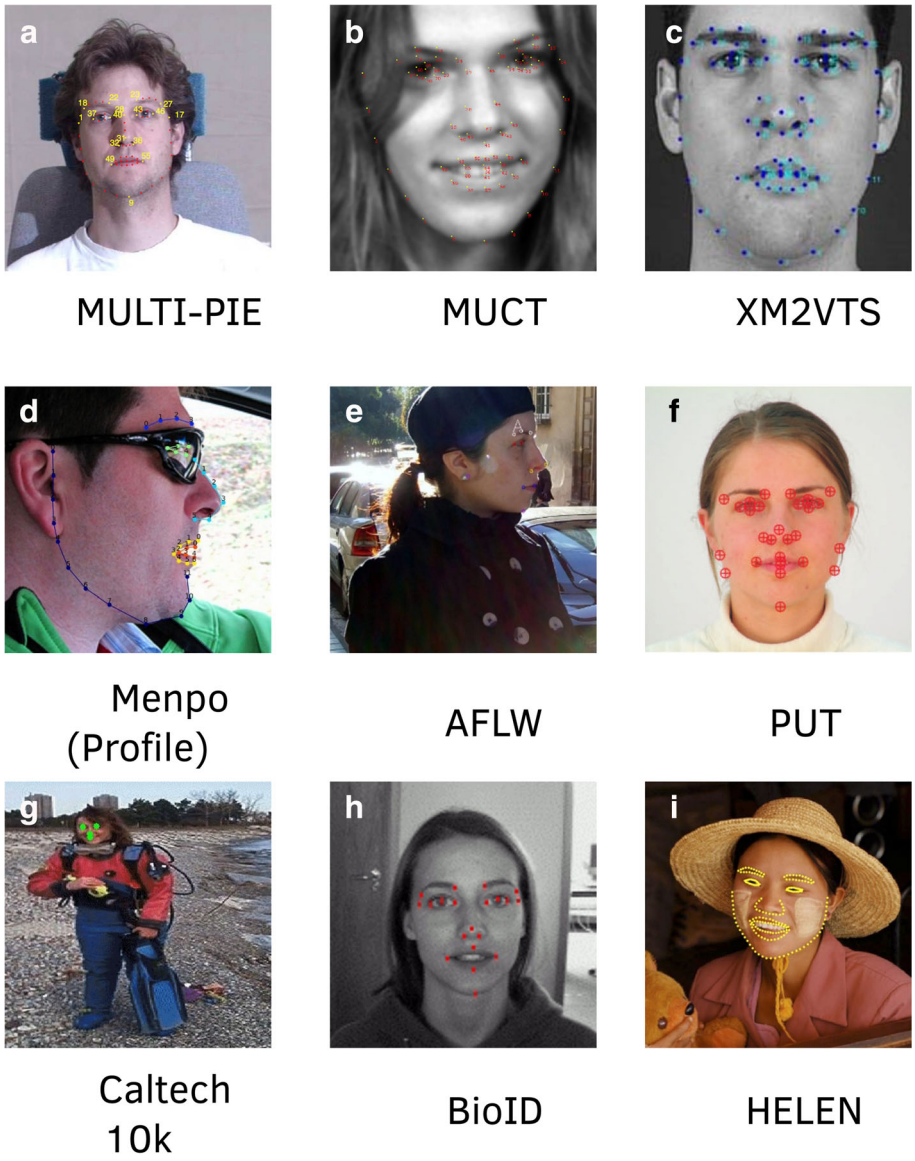

Fig. 2 Examples from public face datasets. These images demonstrate the variety of image types and landmark configurations available within public face datasets. **a** MULTI-PIE [75], **b** MUCT [76], **c** XM2VTS [77], **d** Menpo (Profile) [78], **e** AFLW [79], **f** PUT [80], **g** Caltech 10K [81], **h** BioID provided by BioID AG. ©2001 [82], and **i** HELEN [83]

This problem was identified by Sagonas et al. during the design of the *300W faces in-the-wild challenge* [11]; recognising the variation in annotation schemes between public datasets, Sagonas and colleagues proposed a semi-automatic landmarking tool [12] to provide a single annotation schema (MULTI-PIE/IBUG 68pt) for a number of datasets. This work provided the field with a means of comparing different landmarking methodologies, while utilising the existing datasets. Subsquent to the second *faces in-the-wild challenge* [10], the 300W dataset has been used for performance comparison outside of the context of the landmarking competition [13].

This study reviews and compares the current state of the art in facial landmarking methodologies. To enable this comparison, we require datasets which contain the same landmarking configuration.

> The MULTI-PIE landmark configuration as illustrated in Fig. 3 will be used. This configuration is present within many of the datasets contained within Table 1, including the Menpo and 300W datasets.

2.3.1 Ground truth reliability

With the exception of the annotations provided by IBUG through their semi-automated annotation tool, face

Table 1 Selection of publicly available facial landmarking databases

Name	Images	Subjects	Landmarks	Description	Year
XM2VTS [77]	2 360 (720× 560) RGB	295	68pts[a]	Controlled, 2 head rotation images, 6 images captured during speech	1999
BioID [82]	1 521 (384× 286) Grayscale	23	20	Controlled, front on image, varying expressions	2001
LFW [84]	13 233 (250× 250) mostly RGB	5749 (1680 have≥ 2 images in the set)	MULTI-PIE 68pts[a]	Uncontrolled, originally intended for face identification, images collected from the web	2007
Caltech [81] 10 000 web faces	7 092 varied size mostly RGB	Unknown (10 524 faces in 7 092 images)	4pts	Uncontrolled, images collected from Google images	2007
PUT [80]	9 971 (2048× 1536) RGB	100	30 primary (194 control on subset)	Controlled, portrait images, 5 poses per subject	2008
MULTI-PIE [75]	755 370 (3072× 2048) RGB	337	68pts	Controlled, landmarks for subset, varying expressions	2008
MUCT [76]	3 755 (640× 480) RGB	276	XM2VTS 68pts +4 around eyes (76 total)	Controlled, 5 perspectives 3 lighting conditions Neutral expression or smile	2010
AFLW [79]	2 330 varied size RGB	Unknown	PUT 97pts[a]	Uncontrolled, portrait images, collected from the web (Flickr)	2012
HELEN [83]	2 330 RGB	Unknown	PUT 97pts[a]	Uncontrolled, collected from the web (Flickr)	2012
300W [10, 11]	600 varied size RGB	Unknown	MULTI-PIE 68pts	Uncontrolled, 300 outdoor images 300 indoor images difficult poses and expressions	2013
Menpo benchmark [78]	8 979 varied size RGB	Unknown	Frontal: MULTI-PIE 68 pts Profile: 39pts	Uncontrolled, 6679 frontal images 2300 profile images difficult poses and expressions	2017

[a]The Intelligent Behaviour Understanding Group (iBUG) [71] have made MULTI-PIE 68pts landmarks available for this dataset

datasets require the use of a human annotator(s) to manually identify the fudicial points within the images. This task is critical as any annotation errors will be learnt by the algorithm being trained.

2.3.2 Landmark variability survey

To demonstrate the variability of ground truth landmarks, we performed a study using 20 'workers' on Amazon Mechanical Turk (https://www.mturk.com), who were instructed to identify the MULTI-PIE ground truth landmarks on a single face image. During the study, each annotator used the web-based based 'turkmarker' tool to select the location of each point on the image; ensuring all points were selected, in the correct configuration, and that all results were correctly recorded. In the current literature, ground truth landmarks are typically reported as

Fig. 3 MULTI-PIE landmark configuration. The MULTI-PIE landmark configuration within the MUCT [76] dataset

being annotated by 'experts'. For the purpose of our study the, Mechanical Turk workers were experts.

The turkmarker system is made publicly available for use or modification by the authors through GitLab (https://gitlab.com/docEbrown/turkmarker). A demonstration of the site can also be found via http://benjohnston.info/turkmarker-gh-pages/index.html.

The positions of the ground truth landmarks, identified by the 20 'expert' annotators, are illustrated in Fig. 4. This figure shows that there is increased agreement amongst the annotators regarding the positions of the landmarks on or around the nose, eyes, and the outer border of the lips, compared to the border of the face, eye brows, or joining line of the mouth. It is also interesting to note that while the border of the face was arguably more difficult for the annotators, there is a relative reduction in variability regarding the position of the tip of the chin. These observations are in agreement with Sagonas et al. [10] who performed a similar experiment with three 'expert' annotators and found that it is easier to identify landmarks that lie on distinctive boundaries or junctions such as the corners of the eyes. The raw data and analysis of the results of this survey and the other experimental components of this paper are available online for reference through GitLab.

Given the variation that occurs between different types of landmarks, it is important for achieving accurate predictions that this variation also be considered at the time of dataset selection. Some landmark configuration schemes, such as me17, possess more landmarks at positions with high agreement and if suitable for the problem being solved could improve accuracy. This survey demonstrates the importance of using multiple experts in the ground truth landmarking process to reduce annotator bias.

2.4 Stage 3: regions of interest

The next stage is to extract only those features from the data relevant for solving the defined problem. For automated facial landmarking we must detect and extract the face of interest from the image and discard irrelevant information such as the background. This face detection process forms the first stage of an automated landmarking system and is critical for overall performance. The system must accurately identify and locate the face(s) within the

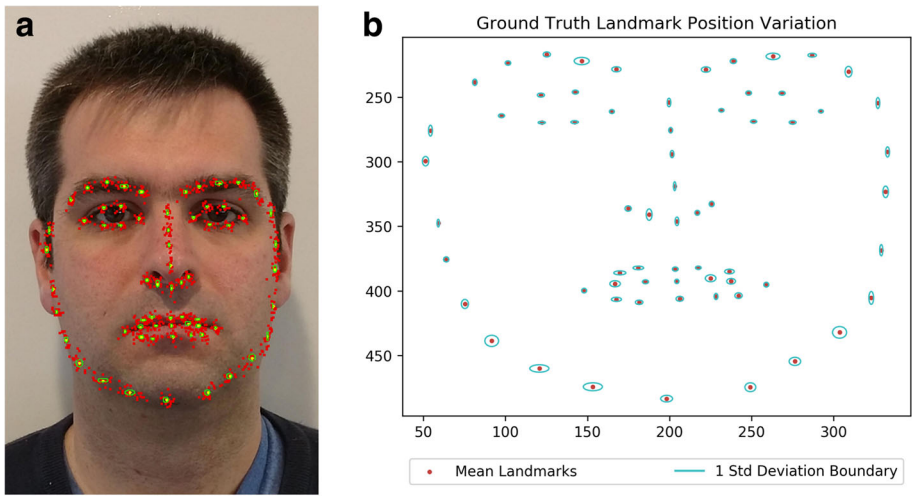

Fig. 4 Ground truth variation. **a** Ground truth landmarks provided by individual annotators. Landmarks in red are the individual positions provided by the annotators, while the yellow points indicate the mean position for each of the landmarks. **b** Ground truth landmarks variation. Points in red are the mean positions for each landmark, while the major and minor axes of the blue ellipses indicate the *x* and *y* variance of the ground truth positions

image, given variations in lighting, pose, expression, and face appearance. All images must first pass through a face detector to extract only the important region(s) of interest (ROI) from the data.

As this article focuses on facial landmarking techniques, we only provide a brief overview of face detectors.

For the purpose of this article:

> We will define the region of interest to be extracted from each image to be a single human face.

Two foundational methods of face detection, which are used as benchmark methods are the Viola-Jones [14] and Histogram of Gradients (HOG) [15] methods. These methods while not currently state of the art achieve reasonable accuracy, processing speed, and are in many 'off-the-shelf' implementations such as in OpenCV [16] and Dlib.

The Viola-Jones method uses a combination of two main techniques: the representation of Haar-like features through construction of an 'integral image' and a series of 'weak' classifiers boosted using the Adaboost learning algorithm. At the very last stage of the cascading classifier is a high-performing perceptron and a decision threshold. Using the boosted classifiers and the perceptron layer Viola-Jones were able to produce a high performing system. While generally not used in more modern applications, Viola-Jones face detection still holds its place in literature as an effective reference method.

Dalal and Triggs [15] proposed the HOG method which produced near-perfect pedestrian detection rates on the MIT pedestrian database. This method converts the image into a series of histograms based on the orientation and magnitude of pixel gradients within the image. An SVM classifier is then used to identify the face within the image based on the values of the histogram. This unique pixel gradient signature has proven to be useful in detecting faces in other image sources such as Li et al. [17] who used camera depth information, as well as classifying if eyes are opened or closed within an image [18]. While HOG may no longer be considered state of the art, the method is still demonstrating its relevance and flexibility through its direct application in hardware. Suleiman and Sze in 2016 [19] were able to acheive object detection within 1080HD video at 60 frames per second while only consuming 45.3 mW of power. Suleiman and Sze propose the use of their hardware in embedded, real-time systems which need to minimise power usage. This hardware based HOG system has a significant power consumption advantage over more modern object detection methods that require the use of a dedicated GPU; as an example, the more recent Nvidia GTX 1080TI GPU consumes 250W of power and recommends a 600 W system power supply, essentially eliminating their use in embedded systems.

First investigated in the early 1990s [20], the use of convolutional neural networks (CNN) in face detection has recently been the focus of much research effort. In 2007, Osadchy et al. [21] demonstrated the performance benefit that can be obtained through the use of CNNs and multi-task learning. Their method trained a CNN (with a similar architecture to the LeNet5 [22]) to map images of faces into a low-dimensional manifold space, parameterised by the pose of the face. In Osadchy's opinion, solving the problem of identifying the face and determining pose are quite related, so that when the two tasks are trained together, an improved performance would be achieved when compared to using separate networks.

The use of a cascade of CNNs was proposed by Haoxiang et al. in 2015 [23] and demonstrated an improvement on the state-of-the-art performance against the Face Detection Data Set and Benchmark (FDDB) [24]. Their method consists of a cascading series of three CNN stages. Each stage is used to detect faces at different resolutions, given potential location windows for a face from the previous stage; where the potential detection window for the first stage is defined as the entire image. Each of the three cascading stages are themselves composed of two separate CNNs, one designed to detect faces within the potential detection window at the given resolution and the following designed to calibrate the bounding boxes of the detected faces to ensure the optimal bounding box for a face is passed onto the next stage.

More recently, an implementation of region-based convolutional neural networks (R-CNNs) claimed to achieve the best performance to date on FDDB amongst published results [25]. Region-based CNNs by Girshick et al. in 2014 [26, 27] apply a CNN to warped 'category independent' regions of an image previously computed by a first stage region proposer such as selective search [28]. The method by Sun and colleagues pre-trained a model using the WIDER-FACE database [24] with hard-negative reinforcement and a feature concatenation stage [29]. Feature concatenation combines the features of selected convolutional layers before they are passed onto the subsequent layers in the network. These extracted features provide greater granularity of the through the combination of high- and low-level convolutional layers.

Other recent studies have investigated the use of CNNs in conjunction with other classification techniques: Zhan et al. in 2015 [30] and Tao et al. in 2016 [31] used CNNs in combination with AdaBoost and SVMs, while Wang et al. in 2016 [32] used a multi-task learning approach combining classification of the presence or absence of a face in addition to the coordinates of the facial bounding box.

For further details on face dection the authors recommend [33, 34].

2.4.1 'Off-the-shelf' face detector performance

This section provides an experimental performance comparison of the OpenCV [16] implementation of the Viola-Jones algorithm as well as the Dlib [35] version of a HOG face detector [15]. This section provides the reader with an expectation of the typical performance that can be obtained using one of these 'generic' face detectors and the value of a specifically trained face detector. The source code for this analysis can be obtained from https://gitlab.com/docEbrown/FacialLandmarkingReview.

2.4.2 Methods

The target face bounding boxes for the BioID, HELEN, MUCT, 300W, and Menpo datasets were computed. The ground truth bounding box was determined by the extremes of the ground truth landmarks for each image, i.e. the minimum and maximum values for x and y. A face was labelled as detected if the bounding box returned by the detector and the ground truth bounding box overlapped by at least 50%

Each of the four pre-trained classifiers of the OpenCV and Viola-Jones algorithm [16] were applied to each of the data sets: OpenCV 1(uses the *haarcascade_frontalface_default.xml*), OpenCV 2 (*haarcascade_frontalface_alt.xml*), OpenCV 3: (*haarcascade_frontalface_alt2.xml*), and OpenCV 4 (*haarcascade_profileface.xml*. The profile face detector (OpenCV 4) was not applied to the BioID and MUCT datasets as these images only contain front-on faces. As per the OpenCV documentation, the default parameters for each detector were used. The Dlib [35] HOG-based face detector was also applied to each of the datasets using default values as per the Dlib documentation.

2.4.3 Results

Tables 2, 3 , and Figs. 5 and 6 summarise the results obtained from applying each of these face detectors to the datasets listed above, with some images in the HELEN, 300W, and Menpo datasets containing multiple faces. Due to the lack of bounding boxes or landmarks for the additional faces, the results count the detected, additional faces in the image which are not the ground truth as false positives.

2.4.4 Discussion

The results of Tables 2 and 3 show that the performance of the off-the-shelf face detectors is highly variable and with the exception of the BioID and MUCT databases is moderate at best. All of the frontal face detectors performed best when applied to the BioID and MUCT datasets (Table 2), as these images were taken in highly controlled environments with consistent lighting, pose, and expression. In contrast, the detectors performed poorly on images from uncontrolled environments, dropping to 67.35% accuracy in the frontal Menpo set. It was also observed that the Dlib detectors outperformed the OpenCV, Viola-Jones-based detectors in all examples, and with fewer false positives. Referring to the profile face detectors (Table 3), all variants performed poorly when applied to the profile Menpo dataset. This can in part be attributed to the significant variance in pose within the dataset with some images being quite similar to those within the frontal Menpo set; however, there is also evidence to suggest that the power of the profile detectors is limited.

2.4.5 Conclusion

In conclusion, the Dlib, HOG-based face detector outperformed all of the OpenCV variants with greater accuracy and fewer false positives. Both the Dlib and OpenCV face detctors performed better when applied to images from controlled environments.

2.5 Stage 4: model definition

After defining the problem, selecting a database, and identifying the faces in the images, the next step is choosing a method for detecting the landmarks. The current literature of facial landmarking methods can be divided into three categories: *generative methods*, *discriminative methods* [36] and methods combining the two, producing *statistical methods* [37]. Generative methods, such as Active Shape and Appearance models as to maximise the probability of facial reconstruction from a deformable

Table 2 'Off-the-shelf' frontal face detector performance

	BioID (N = 1521)		MUCT (N=3755)		HELEN# (N = 2330)		300W# (N = 600)		Menpo (Front)# (N=6679)	
Detector	% Acc	False P	% Acc	False P	% Acc	False P	% Acc	False P	% Acc	False P
OpenCV 1	96.45	184	97.95	616	89.06	3548	75.33	6203	73.24	2775
OpenCV 2	96.12	28	98.24	77	84.64	447	71.17	3574	67.35	387
OpenCV 3	96.45	52	98.62	135	85.88	821	72.00	3966	69.35	670
OpenCV 4	N/A	N/A	N/A	N/A	27.21	274	24.83	1597	22.74	349
Dlib	99.34	1	99.89	6	96.82	351	88.17	2488	87.50	339

#Some images in this dataset contain more faces than just the ground truth. These additional faces may be recorded as a false positive result

Table 3 'Off-the-shelf' profile face detector performance

	Menpo (profile) ($N = 2300$)	
Detector	% Acc	False P
OpenCV 1	15.61	1181
OpenCV 2	6.87	144
OpenCV 3	9.09	266
OpenCV 4	14.87	76
Dlib	24.65	64

[#]Some images in this dataset contain more faces than just the ground truth. These additional faces may be recorded as a false positive result

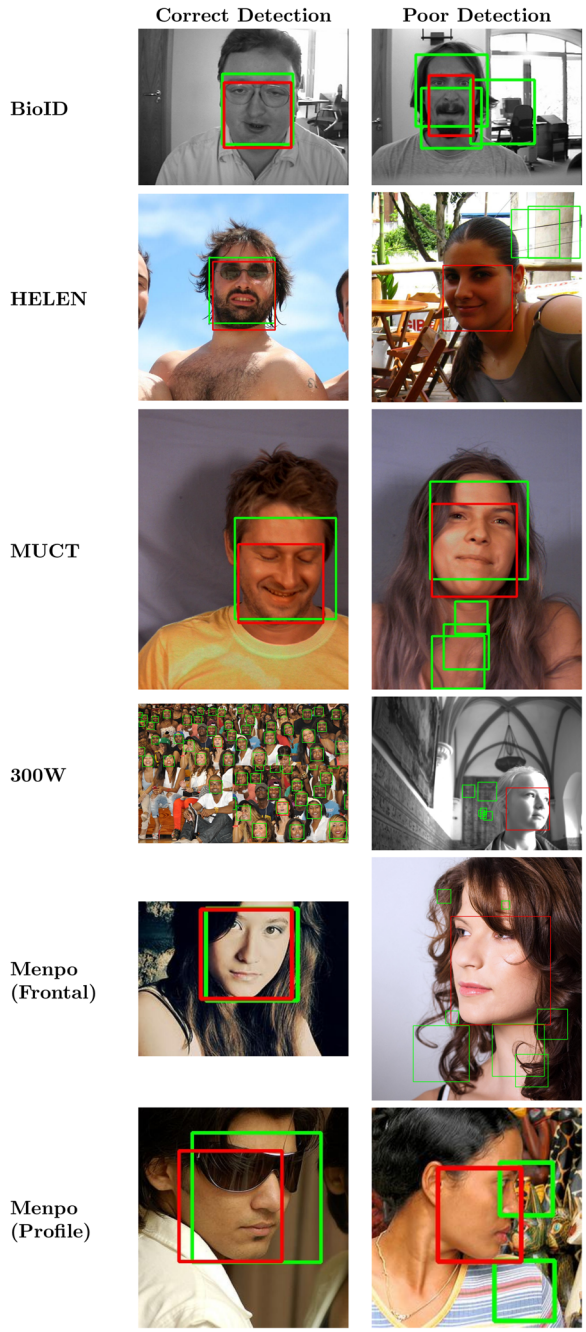

Fig. 5 Face detector results

model. Discriminative methods infer face shape by training a regression function(s) that maps image values to facial landmark coordinates.

Similarly to the face detection process there is a benchmark model to which many other models have been compared. Active appearance models (AAM) [38, 39], an improvement upon Active Shape models (ASM) [40], build a statistical-based representation of the face using the shape information provided by the landmarks and texture information provided by the training images themselves.

We begin our discussion with a description of AAMs and ASMs. These models still play an important role within the field as they provide a benchmark to compare other competing methods [11]. When compared to other techniques (such as deep learning), ASMs and AAMs are not computationally expensive and do not require large sample sizes. They also illustrate the extent the field has progressed over a relatively short period.

2.5.1 Active shape models

The first step in creating an active shape model is to use the Procrustes method [41] to scale and align the landmarks of the training set, while preserving the shape of each training example. After applying Procrustes, a point distribution model (PDM) of the shape is constructed using principal component analysis (PCA) [42] to decompose the data into its constituent components. Having computed and selected the first t principal components for the shape of the ground truth landmarks (P_s), a model is constructed using a vector of weights b_s that can be varied to generate new face shape examples.

In addition to a shape deformable model, a representation of the pixel values around each landmark is required. Typically, this is done by sampling the spatial derivatives of m pixels at either side of each landmark, along the orthogonal vector to the shape contour. When searching for a face within a test image, a pyramid of images is created by scaling and sub-sampling the test image a number of times. The mean shape is then placed at a specified position within the lowest resolution image of the pyramid. This initial placement is very important as it forms the basis of the searching process. The benefit of using an image pyramid is that at lower levels coarse adjustments quickly allows approximation of the correct face location.

For each image pyramid and every landmark in the test shape, a new location is chosen by selecting the pixel along the orthogonal vector to the shape, where the pixel derivatives are most similar to the training set. The parameters

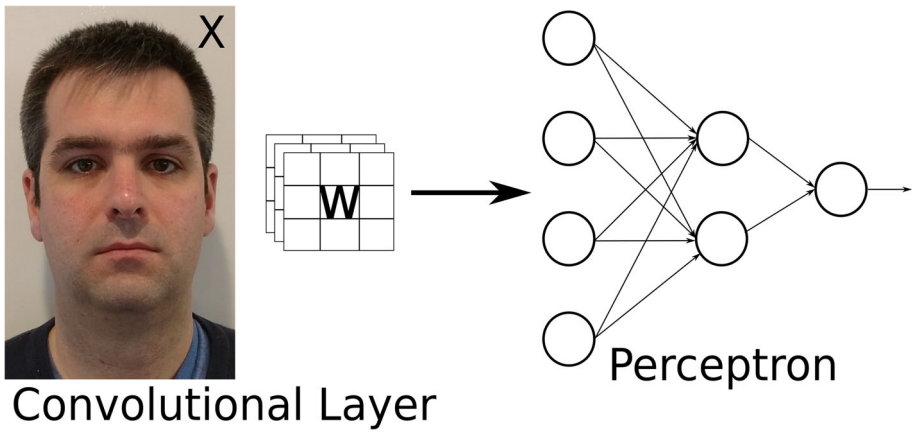

Fig. 6 Simple convolutional neural network. Diagram of a simple convolutional neural network architecture

of the PDM are then updated to match these coordinate locations and the process is repeated, using the updated shape model. This test and update process is repeated until a fixed portion of the predicted landmarks in the test shape stop moving and no more images are available in the pyramid.

2.5.2 Active appearance models

As described above, ASMs do not utilise all of the information that is available within the images, with only pixel information surrounding each landmark being used. AAMs [38, 43] build upon the shape information provided by ASMs by including a detailed texture model of the image. Once the mean shape has been determined a 'shape-free patch' of each training sample is created by warping each training image so the annotated landmarks match the mean shape. Again, PCA is applied to determine a statistical model of texture.

Once statistical models of shape and texture have been generated, they can be combined to form an appearance model that has parameters c to control shape and texture. For a detailed derivation of the combined shape and texture model, see Chapter 5 of Cootes and Taylor's report in [44].

After construction of the combined representation of shape and appearance, the process of searching for the specified facial landmarks within an image is the result of an optimisation problem; aiming to minimise the difference between the grey-level values in the image and the grey-level values described within the combined shape and appearance model.

2.5.3 300W faces in-the-wild challenge

Starting in 2013, the 300W faces in-the-wild challenge [10, 11] has had a significant impact on automatic facial landmark detection research. As discussed in Section 2.3 the 300W competition has provided a benchmark dataset of in-the-wild images with varying lighting, pose, expression, and image location. Due to the influence of this challenge on recent research, our discussion of landmarking literature focuses on the work completed since 2013. For readers interested in literature published prior to 2013, see the review by Çeliktutan et al. [7].

The use of convolutional neural networks and deep learning techniques has dominated recent research in computer vision and particularly in facial landmarking. Given this popularity, it is convenient to discuss convolutional-based and non-convolutional landmarking models separately.

2.5.4 Non-convolutional models

As an entry in the inaugural 300W faces in-the-wild challenge, Baltrusaitis et al. [45] presented a constrained local neural field model. This method built upon the constrained local model (CLM) method described by Cristinacce and Cootes in 2006 [46]; which itself uses the combined shape and appearance model of the AAM. The CLMs, use a series of local patches, one for each landmark, constructed by sampling around the landmarks within the image. The model uses local 'expert' patches around the landmarks with conditional neural fields [47] and a novel non-uniform regularised landmark mean-shift optimisation technique to determine the probability distribution of a landmark location within a patch. This method enabled the landmarking system to learn spatial similarities between landmarks and introduced the requirement of a single landmark to be identified within each patch.

One of the highest performing submissions into the first 300W challenge was the method by Yan et al. [48] which used a series of cascading HOG descriptors to provide a number of landmark estimates for each image followed by a stage that collated the hypotheses from each descriptor into a single result. Given K training samples

$\left\{\mathbf{I_k}, \mathbf{X_k^*}, \mathbf{X_k^0}; k = 1 \ldots K\right\}$ where $\mathbf{I_k}$ is the image/texture information, $\mathbf{X_k^*}$ are the ground truth landmarks, and $\mathbf{X_k^0}$, the initialised landmarks [48] (the mean face shape at the start of the process). The objective is to learn a regression function f minimising the mean squared error between the predicted and ground truth landmarks (Eq. 4). This cascading model learns complex non-linear relationships within the dataset by dividing f into a series of T simpler regression functions ($f_0, f_1, f_1 \ldots f_T$) and using the Hadamard product represented by $\circ$ (Eq. 5). The output of each sub-function is passed as the input to the following stage and a linear transformation $\mathbf{W_t}$ is applied (Eq. 6) to a feature transform $\left(\Phi\left(\mathbf{X_k}^{t-1}, \mathbf{I_k}\right)\right)$ thus encoding the image information around the shape.

$$f = \arg\min_f \sum_{k=1}^{K} \left\| f\left(\mathbf{X_k^0}, \mathbf{I_k}\right) - \mathbf{X_k^*} \right\|_2 \tag{4}$$

$$f = f_0 \circ f_1 \circ f_2 \circ \cdots \circ f_T \tag{5}$$

Where

$$\mathbf{X_k^t} = f_t\left(\mathbf{X_k^{t-1}}, \mathbf{I_k}\right) = \mathbf{W_t} \cdot \Phi\left(\mathbf{X_k^{t-1}}, \mathbf{I_k}\right) \tag{6}$$

The authors noted that their method is highly sensitive to shape initialisation and that the accepted benchmark for face detection of a 50% bounding box overlap was insufficient. To reduce the influence of initialisation, the bounding boxes determined by the face detector are randomly scaled and shifted, generating multiple hypotheses of landmark positions for each image. Given that multiple sets of landmarks are estimated for each image, Yan et al. propose two methods for combining these results. The first method: *learn to rank*, assumes that for each image at least one hypothesis is better than the others and defines a function that ranks and selects the best one from the set. The second method *learn to combine* assumes that for an image the information contained within the entire set of hypotheses is complementary to the final solution and thus defines a function that combines the hypotheses into a single result. The parameters of both of these functions can be determined by solving with a structural SVM.

2.5.5 Cascade shape regression models

Compared with their generative counterparts, discriminative landmarking methods have demonstrated superior performance in uncontrolled conditions [36]. This performance has been further improved through the development of a cascade of linear regression models that are capable of describing complex non-linear relationships. Asthana et al. [36] outline such a method which has resemblance to AAMs models, defining a shape model as

$$\mathbf{x}(\mathbf{p}) = s\mathbf{R}\left(\bar{\mathbf{x}} + \Phi_\mathbf{s}\mathbf{g}\right) + \mathbf{t}; \qquad i = 1 \ldots N \tag{7}$$

Where $\mathbf{x} \in \Re^{2N \times 1}$ is a vector form of $\mathbf{X} \in \Re^{N \times 2}$, s is the scale, $\mathbf{R}$ is the rotation matri, and $\mathbf{t}$ is a translational vector. The vector $\Phi_\mathbf{s}\mathbf{g}$ specifies non-rigid shape variation; thus, $\mathbf{p} = \left[s; r_x; r_y; r_z; t_x; t_y; \mathbf{g}\right]^T$. By specifying a set of shape parameters $\mathbf{P^*} = \left\{\mathbf{p_i^*}\right\}_{k=0}^{K}$ that are defined by the ground truth shape. Asthana et al. define their objective as learning a function from an initial estimate of $\mathbf{p}$ which produces the ground truth $\mathbf{p}^*$. The function that converts the initial shape estimate is a sequence of regression functions [49]. For each of the K training shapes, the parameters defining the shape model are randomly sampled within a defined range around the ground truth parameters $\mathbf{P^*}$ producing a set of J perturbed shape parameters $\left\{\mathbf{p_j^{(1)}}\right\}_{j=1}^{J}$. A linear relationship between the input image $\mathbf{I}$ and the perturbed parameters $\mathbf{p_j^{(1)}}$ is described as

$$\mathbf{p}^* = \mathbf{p^{(1)}} + f\left(\mathbf{I}, \mathbf{x}\left(\mathbf{p}^{(1)}\right)\right) W + b \tag{8}$$

where the function f could return SIFT [50] or HOG features around each landmark matrix. The linear relationship described by Eq. 8 is unable to map the perturbed shape parameters $\mathbf{p^{(1)}}$ to the ground truth $\mathbf{p}^*$. Thus, a cascading series of regression functions is trained by finding $\tilde{\mathbf{W}} = [W; b]$ to solve the following problem:

$$\arg\min_{\mathbf{W^{(1)}}, \mathbf{b^{(1)}}} \sum_{k=1}^{K} \sum_{j=1}^{J} \left\| \mathbf{\Delta p_{kj}^{(1)}} - \tilde{\mathbf{f}}\left(I_i, p_{kj}^{(1)}\right) \tilde{\mathbf{W}}^{(1)} \right\|^2 \tag{9}$$

Where $\tilde{\mathbf{f}}(\mathbf{I}, \mathbf{p}) = [f(\mathbf{I}, \mathbf{p}) \quad 1]$ and $\Delta\mathbf{p_{kj}^{(1)}} = \mathbf{p_k^*} - \mathbf{p_{kj}^{(1)}}$ and j is the index of perturbation. By repeatedly solving for $\tilde{\mathbf{W}}^{(1)}$ and updating $\mathbf{p_{kj}}$, the perturbed parameters approach the ground truth. In their work, Asthana et al. propose a parallel method of cascade linear regression that does not rely on the perturbations of previous iterations and benefits from executing the process in parallel. This *parallel cascade of linear regression* uses only the statistics of the previous level, removing the need to propagate through all of the samples and iterations. One further benefit of this method is that if additional training data is made available, the costly process of retraining through all combinations of iterations and perturbations is not required as the cascade of regression functions are trained individually.

Deng et al. in 2015 [37] extended the cascade shape model, describing a multi-view, multi-scale, and multi-component cascade shape model (M^3CSR). Deng's method first applied a six-component deformable part model face detector to divide the training samples into front, left, and right profile views. By grouping the images, the variation within each view set is reduced, improving the model's performance with an increased range in pose.

After finding the HOG descriptors for each of the images in each set, the cascade shape regression process is executed in a coarse to fine manner; starting with a small instance of the face and then doubling the face's size to produce a second scaled image. This coarse to fine process improves the speed and accuracy of convergence by allowing for gross changes to landmark position at the coarse level and precision adjustment within the higher resolution image.

The final aspect of the M^3CSR process is the multi-component stage to compensate for differences in landmark stability, similar to that described in the Section 2.3.2. Given differences in facial expression or natural variation in the location of particular landmarks such as the tip of the nose compared to the cheek, an additional alignment process was executed for each landmark:

$$\arg\min_{W_t} \sum_{k=1}^{K} \left\| \left(\mathbf{X}_{\mathbf{k}}^{*} - \mathbf{X}_{k}^{\mathbf{t}-1} \right) - \mathbf{W}_{\mathbf{t}} \Phi \left(\mathbf{I}_{\mathbf{k}}, \mathbf{X}_{\mathbf{k}}^{\mathbf{t}-1} \right) \right\|_2^2 \quad (10)$$

where t is the iteration number, $\mathbf{X}_{\mathbf{k}}^{\mathbf{0}}$ is the initialised shape coordinates, $\mathbf{W}_{\mathbf{t}}$ is the linear transform matrix, and Φ represents the shape index feature descriptor, corresponding to either the front, left, or right profile views. Note the similarities between Eqs. 9 and 10 which follows the similar processes used by Deng et al. and Asthana et al.; applying the variation within the landmarks themselves, instead of intentional perturbations to construct an aligned shape model.

Having separated the training images on the basis of pose and doubled the scale of the images, the training process comprises of iterating through each set of poses and scales: generating 10 different shape initialisations, computing the HOG features around each landmark, finding $\mathbf{W}$, and updating the shape $\mathbf{X}_{\mathbf{k}}^{\mathbf{t}} = \mathbf{X}_{\mathbf{k}}^{\mathbf{t}-1} + \mathbf{W}_{\mathbf{t}} \Phi \left(\mathbf{X}_{\mathbf{k}}^{\mathbf{t}-1}, \mathbf{I}_{\mathbf{k}} \right)$. During their experimentation, Deng and colleagues also confirmed the performance increase of M^3CSR when compared with M^1CSR (multi-view cascade shape regression) and M^2CSR (multi-view, multi-scale cascade shape regression) models.

2.5.6 Convolutional neural network landmarking models

The use of CNNs has gained much popularity recently within the field. This popularity can partly be attributed to the availability of large datasets and high performance, affordable computational hardware such as graphics processing units (GPUs). The use of CNNs have also increased after the significant increase in ImageNet classification performance achieved by Krizhevsky et al. [51] with AlexNet. Many of the current leading classifiers on benchmark datasets such as MNIST [22, 52], CIFAR-10 [53, 54], and CIFAR-100 [53, 55] are based on CNNs. This has been particularly evident in computer vision where CNNs or deep learning has provided effective solutions for face detection, facial landmark prediction, and other problems.

2.5.7 Convolutional neural networks

In its most basic form, a convolutional neural network is a simple perceptron (or neural network) with a preceeding stage that performs a convolution operation ($\mathbf{Q} = \mathbf{I} * \mathbf{W} + \mathbf{b}$) on an input image $\mathbf{I}$, given a set of weights $\mathbf{W}$ and biases $\mathbf{b}$ [56]. Similarly to the weights within the perceptron, backpropagation is applied to optimise $\mathbf{W}, \mathbf{b}$ for the given cost function to adjust the weights according to their contribution to the error. Thus, the convolutional layers form a set of feature extracters for the system. This is one of the most powerful aspects of CNNs; rather than using manually crafted features such as Haar-like or HOG features, the system automatically 'learns' the optimal features for the dataset and the objective of the network. Additionally, CNNs, unlike many other model designs, are able to continue to improve their performance as data is added to the training set.

As outlined in the Section 2.5.4 as well as in [11] and [10], the first two 300W competitions received submissions using a number of different methodologies including deep learning. In contrast to the 2017 Menpo Facial Landmark Localisation Challenge [57], all submissions to the competition utilised deep learning. Deep learning methods have won every recent facial landmarking challenge: Zhou et al. in 2013 [58], Fan and Zhou in 2015 [59], and Yang et al. [60] winning both frontal and profile competitions in 2017. The high performance of these techniques prompted the organisers of the Menpo Challenge to ask "is the current achieved performance good enough?" [61].

One of the most common designs of deep learning models in facial landmarking is the cascading structure where a number of different convolutional neural network stages are connected sequentially to produce the final landmark predictions. Typically, cascading networks are employed in a coarse to fine manner, where earlier networks in the cascade make more gross predictions regarding landmark position and later stages make fine positional adjustments. An early user of this methodology was Sun et al. in 2013 [62]. They used three cascading levels to predict five landmarks on the face (see Fig. 7). After applying the face detector, Sun et al. constructed three CNNs within the first stage: the first CNN received an image of the entire facial region, the second an image of the eyes and nose, and the third of the nose and mouth. The first stage estimated the approximate landmark locations, thus each network was trained to make coarse predictions of the landmark positions within the corresponding region of the image. The three CNNs observed each landmark at least twice, and thus multiple predictions for each landmark were produced. The average of each of these

Fig. 7 5 point landmark configuration. The landmark configuration used by Sun et al. [62]

predictions was calculated and used to define local image patches which were provided to the second convolutional stage. As can be seen in Fig. 8, the design of the second and third convolutional stages is somewhat similar; both being provided with localised image patches on which to train and thus both being restricted to making small adjustments on the previous predictions. Similar to the first convolutional stage, multiple predictions were made for each landmark by the second and third stages and thus the averages were again computed. The CNNs within each stage used repeating, alternating convolutional and pooling layers followed by two fully connected layers; for more details see [62].

Zhou et al. and Fan and Zhou also employed a cascading coarse to fine process in their work in 2013 and 2016 [58, 59], taking a slightly different approach in each. The earlier work employed a method similar to that of Sun et al. [62]; defining a 3-stage cascading CNN structure where the first stage made predictions at 'face-level' while the latter two made predictions with localised patches [58]. After detecting the bounding box of a face within an image, Zhou et al. extracted two sub-images of the face with one centered on the 300W inner facial landmarks and the other centered on the 300W outer facial landmarks. These sub-images were provided to the first stage which used two distinct CNNs to separate predictions for inner and outer facial landmarks. Similarly to Sun et al., following the coarse prediction stages, localised patches (in this case of the eyes, eyesbrows, nose, and mouth) were used in the following two CNN stages to make precise position adjustments. In their paper, Zhou notes that due to time constraints, the final predictions made for the outer landmarks were purely those provided by the first CNN stage; there is no technical reason why the outer landmark set could not be passed through similar subsequent stages to potentially improve performance. Each of the CNNs within the structure utilised three convolutional layers, each followed by max-pooling layers. An interesting aspect of their design is the inclusion of an unshared convolutional or local-receptive field layer prior to a fully connected layer. Unlike the typical convolutional layer which shares the same weights as it convolves across all of the input, unshared convolutional layers use different weights within the convolutional kernel at locations of the input. Shared weight convolutional layers are known to be useful in extracting features and removing translational variance within images [56]; the use of unshared weights allows for translational variance to be included into the convolutional kernels. In scenarios where the general structure of the image is consistently positioned, such as a centered face, the use of unshared weights can allow the network to learn more subtle features within the known structures such as the eyes, eyebrows, and mouth.

In 2016, Fan and Zhou claimed near human landmarking performance with a 2-stage cascade of deep CNNs [59] similar to that of Sun et al. but with an additional alignment stage following the first level of predictions. This alignment process, similar to that used in Procrustes analysis [41], transforms the original predictions and input images into a common scale and rotation space prior to dividing the input image into localised patches. For each landmark, a separate transformed patch was provided to an individually trained CNN to refine the initial predictions and produce the overall result. Unlike in [58], Fan and Zhou used a more conventional CNN design in this model, alternating shared weight convolutional layers and pooling layers followed by two fully connected layers.

This method of combining cascade CNN stages with image and landmark alignment was extended by Chen et al. [63]. This updated method executed the alignment process prior to the stage 1 and included an additional refinement stage following the component level stage 2; with the third stage performing alignment at the individual landmark level. With regards to the composition of the CNNs themselves, Chen and colleagues implemented a first stage network using skip-connections and channel wise convolutions.

When compared to previous years, one could argue that recent facial landmarking publications have implemented CNNs and deep learning techniques with more variety and creativity than before. As the effects of various cascading structures of more 'conventional' CNNs have been thoroughly explored, recent studies have investigated dramatically different network designs and methodologies. Similarly to the multi-task face detection techniques described previously, Zhang et al. [64] supplemented the image and landmark information with auxiliary attributes such as gender, pose, facial expression, and whether subjects are wearing glasses. By training a deep CNN with a sparse set of landmarks and the auxiliary features, Zhang

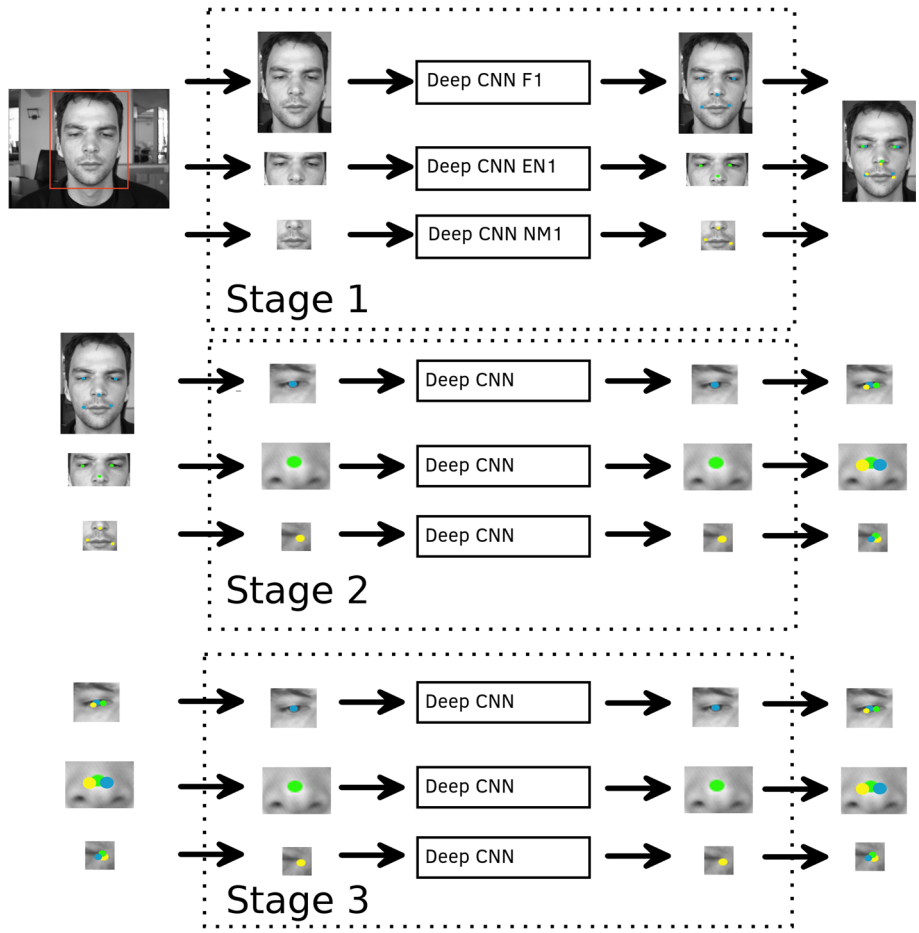

Fig. 8 Cascading convolutional networks. The cascading CNN structure used by Sun et al. [62]

et al. were able to demonstrate an increase in performance when transferring the model to dense landmark predictions.

Yang et al. [60] won both the frontal and profile competitions of the 2017 Menpo landmarking challenge using a *stacked hourglass* convolutional network architecture. The conventional CNN design increases the depth of each layer throughout the network, while decreasing the height and width, until the last layer produces a very deep, but small output. The stacked hourglass approach, first proposed by Newall et al. in 2016 [65], differs in that after each max-pooling stage a copy of the convolved output is taken (Fig. 9). Once the network has reached its lowest resolution, the output is passed through a series of upscaling methods, combining the previous copy of the output at each stage which allows for input at a variety of scales and produces a symmetric topology. The coined name stacked hourglass is applied to this method as Newall stacked a number of these symmetric networks together to perform the final prediction. Following a supervised alignment process using Procrustes analysis, Yang et al. applied four stacked hourglass networks to face images of 256×256 pixels to achieve state of the art results.

Wu and Yang's approach to facial landmarking [66] investigated the variance and bias with regards to facial pose, expression, occlusion, and other factors within and between each of the benchmark datasets. Their analysis hypothesised that the variability between the datasets could lead to overfitting for particular factors and thus poor generalisation, for example, in the 300W data set 14.16% of the faces have a scream expression while in HELEN, none of them do [66]. Given this hypothesis, the landmarking model described by Wu and Yang comprises of two individual CNNs; the first, known as the *Dataset-Across Network (DA-Net)* is a more conventional CNN except that the last fully connected layers are split into n layers where n is the number of datasets being used to train the network. *DA-Net*, by splitting the final fully connect layers, attempts to learn the general features that comprise faces within each of the datasets, e.g. shape and landmark relationships. It is important to note that this network does not produce multiple landmark

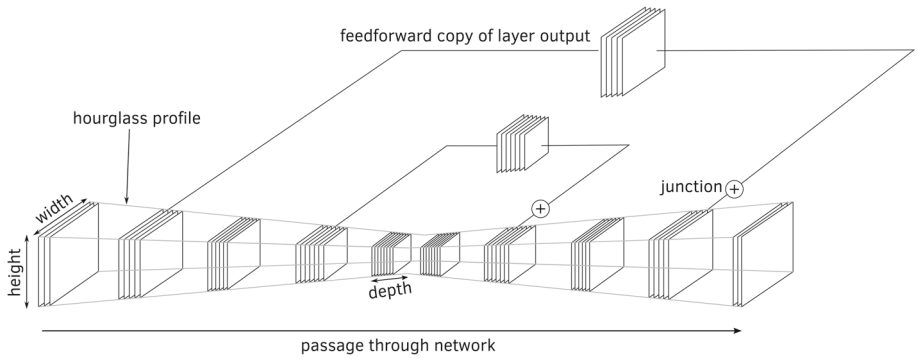

Fig. 9 Example of a stacked hourglass CNN. The stacked hourglass CNN not only follows a conventional CNN structure by reducing image width and height, and increasing depth but is also followed by an inverting structure which returns the width, height, and depth to original or near-original values

estimates for each image but rather produces estimates for an image using the fully connected layers corresponding with its appropriate dataset of origin. This model is essentially training individual regressors for each dataset but shares a common set of feature extractors via the preceding convolutional stages.

Wu and Yang acknowledge that *DA-Net* would still suffer from intra dataset bias and proposed a *Candidate Decision Network (CD-Net)* to address these biases; e.g. Wu and Yang attempted to account for the bias of views within the Menpo profile dataset, with 72.9% of the images right viewed. Using a combination of left view samples and right view samples Wu and Yang trained a DA-Net and CD-Net to decide whether landmark predictions made by DA-Net are more suited to an image or its corresponding left-right flipped image. In this way, CD-Net essentially forms a decision layer whereby two versions of an image with DA-Net predictions are provided and the best result selected by CD-Net. Wu and Yang note that CD-Net is provided to mitigate bias, when applied to new datasets performance may vary, depending on the variance within the dataset. In other examples, such as varying facial expression an augmentation process would need to significantly alter the image to provide DA-Net and CD-Net which could further increase error or bias inthe result.

One of the most interesting aspects of recent deep learning and CNN work has been the merging of more classical generative modelling techniques with that of the modern deep methods. He et al. [67] describe a fully end-to-end cascaded convolutional neural network (FEC-CNN), whereby each layer in the cascade builds upon the results of previous predictions. Each FEC-CNN (H) is composed of individual CNNs (C_t), which are trained using image patches extracted using a function Θ, given the landmark locations and the results of the previous layer. For the first sub CNN C_0, the mean shape $\bar{\mathbf{X}}$ is provided to determine the image patches and C_0 is trained to predict the landmark positions. The difference between the target landmarks and the predicted $\Delta\mathbf{X_t}$ are then added to the input shape ($\mathbf{X_0}$) to produce the image patches for the next sub CNN. This process is then repeated for T stages of sub CNN where the final predicted shape $\mathbf{X}$ is given by

$$\mathbf{X} = H(\mathbf{I}) = \sum_{t=1}^{T} \Delta\mathbf{X_t} + \mathbf{X_0} \tag{11}$$

$$\Delta\mathbf{X_t} = C_t(\Theta(\mathbf{I}, \mathbf{X_{t-1}})), t = 1, \ldots, T \tag{12}$$

What is particularly interesting about this method is that as each sub CNN takes patches from the original image and the output is fed into the next layer, the errors can be backpropagated throughout all CNNs within the network, leading to a significant performance improvement over networks which are trained in isolation. This process of using the errors of the previous sub CNN is similar to the process of constructing an active shape model, i.e. starting with the mean shape, determining the corresponding errors, updating the model, and repeating. He and colleagues successfully applied this FEC-CNN model in [68] with a preceding CNN-based face detector which reduced landmark sensitivity due to bounding box initialisation by providing the minimum enclosing rectangle.

Zadeh et al. proposed a *convolutional experts constrained local model* (CE-CLM) [61] which sought to blend the feature extraction ability of CNNs with the landmark specificity of CLMs [69]. They believed that the CLMs are currently under-performing alternative models due to a restricted ability to accommodate the large image variations about each landmark and thus introduced *expert local detectors* capable of modelling such variation. As the CE-CLM algorithm is based upon a CLM, an iterative parameter update is at the core of the process and aims to optimise

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \left[\sum_{i=1}^{N} -D_i(\mathbf{x_i}; \mathbf{I}) + \Lambda(\mathbf{p}) \right] \tag{13}$$

where $\mathbf{x_i} \in \Re^{1\times 2}$ is a subset of $\mathbf{X} \in \Re^{N\times 2}$. Similarly to Eq. 8, $\mathbf{p}^*$ are the optimal parameters defining landmark position and $\mathbf{p}$ denotes the current estimate for the location of $\mathbf{x_i}$. $D_i(\mathbf{x_i}; \mathbf{I})$ is the probability of the ith landmark being in position $\mathbf{x_i}$ for input image $\mathbf{I}$ and $\Lambda(\mathbf{p})$ is a regularisation function. One should note that the definition of $\mathbf{p}^*$ and $\mathbf{p}$ are identical to that described within Eq. 7 and in contrast to many other methods Eqs. 7 and 13 use ground truth and predicted landmarks in vector form $\mathbf{x}$ not the corresponding matrix form $\mathbf{X}$. The most important part of the CE-CLM design is the *Convolutional Experts Network (CEN)* which takes landmark localised image patches and returns the probability map $D_i(\mathbf{x_i}; \mathbf{I})$, which as shown in Eq. 13 is crucial in finding the optimum model parameters.

The CEN is a specially crafted network composed of three distinct layers: a contrast normalising convolutional layer, a generic convolutional layer, and a mixture of experts (ME) convolutional layer. Upon providing the network with a localised image patch, the contrast normalising layer performs Z score normalisation on the image prior to convolving with a ReLU-based kernel of the generic convolution layer. The first two layers of the network essentially prepare the data for the third and final ME-layer which produces the final alignment probabilities for the image patch. The ME-layer is a convolutional layer with a kernel using a sigmoid activation function which produces the response map or probabilities of landmark alignment within an image patch. The response maps are then combined with a non-negative weight final layer which too uses a sigmoid activation function to compute $D_i(\mathbf{x_i}; \mathbf{I})$. Zadeh et al. noted that a 1×1 kernel size with no pooling layers was selected to increase the resolution of the landmark prediction space and that the ME-layer had a significant impact on the overall performance. Unlike other CNNs, increasing the depth of the CEN did not improve the performance of the network while changes to the ME-layer such as removing the constraint of non-negative weights result in a significant reduction in performance.

Consistent with CLMs and cascade regression models, Zadeh et al. used point distribution models (PDMs) to control landmark locations and penalise irregular shapes through $\Lambda(\mathbf{p})$. Using Eq. 7 and given an initial CE-CLM estimate $\mathbf{p}$, the required updates $\Delta\mathbf{p}$ are iteratively applied in response to the positions provided by CEN to determine the final model that maps the input image to the corresponding face shape.

The final CNN model discussed is the *Deep Alignment Network (DAN)* by Kowalski et al. [70], which was inspired by the cascade shape regression (CSR) model. Like CSR models, DAN starts with the mean shape $\bar{\mathbf{X}}$ as the initial shape estimate $\mathbf{X_0}$, which is refined by a single stage of a deep neural network representing an iteration of a CSR. In addition to being inspired by the CSR methodology, this method also possesses components similar to that described by He et al. [67] and Zadeh et al. [61]. Each stage of DAN (see Fig. 10), with the exception of the first, is comprised of a CNN landmark prediction process followed by a landmark transformation step and a connection layer which passes the information onto subsequent DAN stages. The transformation layer is comprised of multiple sub-layers, each which produce one of five inputs for the subsequent DAN stage.

The first component of a DAN stage is a four-layer CNN with pooling and two fully connected layers, trained to estimate the position of each of the landmarks ($\Upsilon_\mathbf{t}$). For stages $t = 2, \ldots T$, the CNN is provided with a warped image $\mathbf{W_t}$, a landmark response map $\mathbf{E_t}$, and a feature image $\mathbf{L_t}$ from the previous stage; for the first stage ($t = 1$), only the original image $\mathbf{I}$ is provided. The CNN predictions ($\Upsilon_\mathbf{t}$) are added with the transformed shape provided by $\Gamma_t(\mathbf{X_{t-1}})$.

The second step in a DAN stage is landmark transformation which is used to warp the input image $\mathbf{I}$ and the current landmark predictions $\mathbf{X_t}$ to the canonical or mean shape $\bar{\mathbf{X}}$ and to construct $\mathbf{W_t}$. In the described implementation, Kowalski et al. used an affine transform to warp the image and landmarks. They note that other transforms can be used provided that the transform Γ_t is invertible as the output of every stage must be able to be returned into the original image space.

$$\mathbf{X_1} = \mathbf{X_0} + \Upsilon_\mathbf{1} \tag{14}$$

$$\mathbf{X_t} = \Gamma_t^{-1}(\Gamma_t(\mathbf{X_{t-1}}) + \Upsilon_\mathbf{t}) \tag{15}$$

$$\mathbf{X_{DAN}} = \sum_{t=1}^{T} (\Gamma_t(\mathbf{X_{t-1}}) + \Upsilon_\mathbf{t}) \tag{16}$$

Following the transformation step, a landmark response map is generated where the highest intensity values indicate predicted landmark locations. The response map is generated using the landmark estimates provided by the previous stage, thus propagating the predictions throughout the network.

Complementing the response map is the feature image layer $\mathbf{L_t}$, the last output of a DAN stage. $\mathbf{L_t}$ is an image generated from the output of the first fully connected layer of the previous CNN ($t-1$). The output of the dense layer is reshaped to a 2D layer and is provided to the next stage ($t+1$).

Having defined and generated all of the required inputs for the next DAN stage, T stages can be concatenated to form the overall model. While this approach resembles

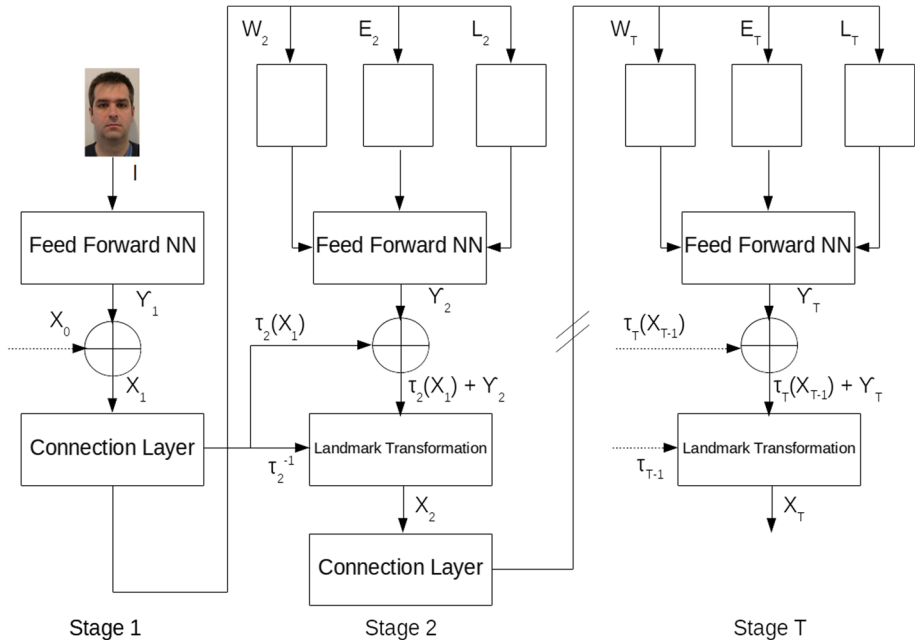

Fig. 10 DAN structure

the FEC-CNN described by He et al. [67], the training methodology differed considerably. He and colleagues observed a significant increase in performance by training using an end-to-end approach and backpropagating the errors throughout each sub-CNN. Kowalski and colleagues however recorded a reduction in performance using this end-to-end approach and as such trained each DAN stage in isolation.

2.5.8 Benchmark model performance

We now examine the performance of menpofit, a Python-based open-source implementation of active appearance models. This section provides the reader with the 'typical' performance that can be achieved using AAMs. The experimental results are reported alongside results published using deep learning methods; allowing for an indirect comparison. The source code for this section can be obtained from GitLab.

2.5.9 Methods

The BioID, HELEN, MUCT, 300W, and Menpo datasets were used. For the HELEN, MUCT, 300W, and Menpo (frontal) sets the complete MULTI-PIE 68 landmark configuration was used as provided in either the original dataset or the ibug coordinates [71]; for the BioID and profile Menpo datasets, the provided landmark configurations were applied.

To eliminate the potential for additional error to be introduced by a face detector algorithm, the face for each image was first extracted using a bounding box 10% larger than the extremes of the ground truth landmarks. Once each face was extracted, each data set was randomly split into training and test subsets with 70% of the data allocated for training. Each of the training sets were then used with the MenpoFit software to train a holistic active appearance model (hAAM) which uses the entire image to construct the model and a patch active appearance model (pAAM) which only uses patches of the image surrounding the individual landmarks. Each of the models were trained using a two-level image scaling pyramid where the scaling factors were set to 0.5 and 1. Due to the large sample size of the HELEN, MUCT, and Menpo datasets, training was completed for these sets using mini-batches where each batch comprised 256 samples. The BioID and 300W (indoor, outdoor, and combined) datasets were trained without the use of mini-batches.

For each image in the test set, landmark predictions were made by the specifically trained models and the normalised root mean squared error was calculated to compare performance. For the BioID, HELEN, MUCT, 300W, and frontal Menpo datasets, the Euclidean distance between the outer corners of the left and right eyes were used for normalisation; in the MULTI-PIE configuration, this was the distance between landmarks 37 and 46, for BioID landmarks 9 and 12 were used. For the profile Menpo datasets where the corners of both eyes were not present, the diagonal length of the face as defined by the distance between landmarks 10 and 13 was used as the normalising factor.

For all methods and all datasets, cumulative error distribution curves were produced. For the BioID, 300W, and the Menpo datasets, error curves indicating state of the art performance were overlayed and referenced.

All models were trained and tested using the high performance computing facilities provided by the Sydney Informatics Hub at the University of Sydney which include Dell PowerEdge R630 and C6320 Server nodes. Each PowerEdge R630 node possess 24 Haswell Intel Xeon E5-2680 V3 cores with 128 GB of DDR3 RAM, while the C6320 nodes possess 32 Broadwell Intel Xeon E5-2697A-V4 cores also with 128 GB of DDR3 RAM. The resource utilisation of each model was recorded and is provided in the Section 2.5.10.

2.5.10 Results

Figures 11, 12, 13, 14, 15, 16, 17, and 18 describe the performance of each method on each dataset using cumulative error distribution curves, while Table 4 indicates the resources required to produce and test each model. Note that CPU utilisation is recorded as the average number of CPUs used over the course of executing the script.

2.5.11 Discussion

Observing the AAM curves across the figures, it can be seen that patch-based AAMs outperform holistic models for almost all datasets; with near-identical performance using HELEN (see Fig. 15). This is particularly evident in the profile Menpo results (Fig. 18) with the holistic method essentially unable to model the data. As noted by [66] and [61], the profile Menpo dataset is particularly challenging due to the variety of facial poses (some could plausibly be in the frontal dataset) and an imbalance of left/right side facing profile images. The failure of the holistic model can be attributed to this variance and the inability of the model to represent it. Conversely, the patch-based model, while including all of the shape information, only uses localised image information around the areas of interest and thus ignores some of the irrelevant information.

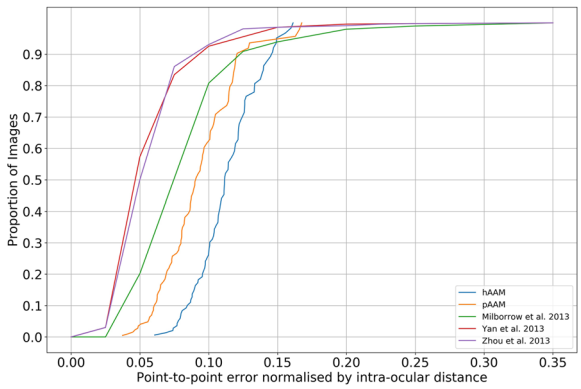

Fig. 12 300W (indoor) cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Yan et al. [48], Zhou et al. [58], and Milborrow et al. [72]

Comparing these 'typical' AAM performance metrics to published literature, it can be seen that the patch-based AAM model for the BioID dataset (see Fig. 11) is comparable to the AAM as described by Cristinacce and Cootes [46]. Both CED curves follow a similar path; however, the maximum error produced by the patch-based AAM model is lower than that reported by Cristinacce and Cootes. When reviewing the cumulative error distribution curves in comparison to more recent deep learning techniques, it is not appropriate to make a direct comparison of performance between the hAAM, pAAM, and reported methods. The models submitted to the 300W competition were trained using several different datasets including the 300W training set prior to testing on the

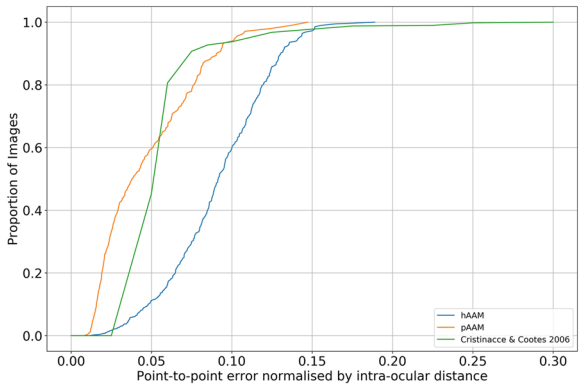

Fig. 11 BioID cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Cristinacce and Cootes [46]

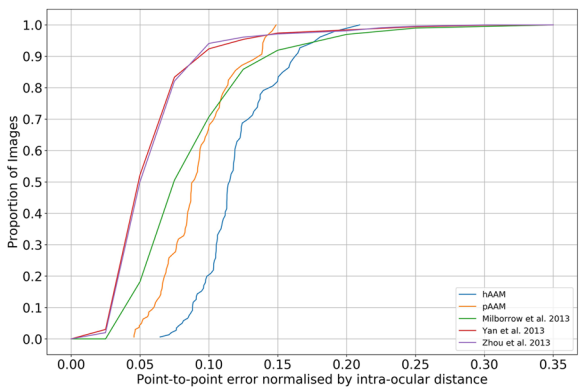

Fig. 13 300W (outdoor) cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Yan et al. [48], Zhou et al. [58], and Milborrow et al. [72]

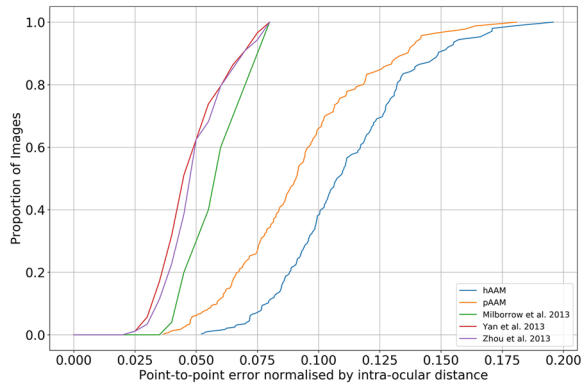

Fig. 14 300W (combined) cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Yan et al. [48], Zhou et al. [58], and Milborrow et al. [72]

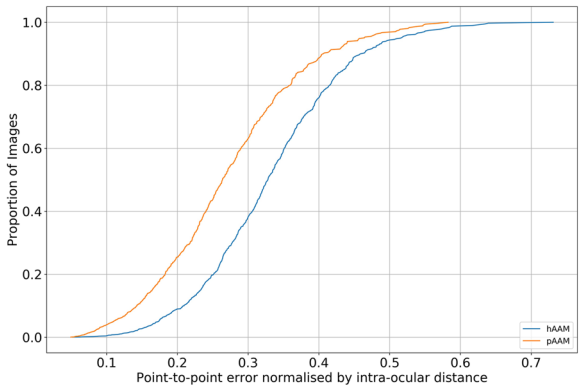

Fig. 16 MUCT cumulative error distributions. Comparing the performance of holistic AAMs and patch AAMs

300W test set. The training data for our AAM models and the literature differ significantly, thus preventing a direct comparison with Figs. 12 and 13. Despite this, the ASM competition entry by Milborrow et al. [72] does enable a direct comparison between classical generative and more recent methods. Zhou et al. [58] using a coarse to fine convolutional network cascade and Yan et al. [48] using a combination of multiple regressors produced an elevated performance on both the 300W indoor and outdoor datasets. It is also possible to compare the performance of Zhou et al.'s convolutional model to the higher performing fully end-to-end CNN described by He et al. [67] as the authors reported performance against the combined 300W dataset (see Fig. 14).

With regards to the frontal and profile Menpo datasets while indicative performance of hAAMs and pAAMs are provided, no meaningful comparison can be made to the reported performance in the literature. The training and test sets are simply too different. What is interesting to note is that reviewing Figs. 17 and 18, in addition to the Menpo competition results [57] is that many of the entries used some CNN-based method. The highest performers employed sophisticated deep learning approaches such as stacked hourglass and fully end-to-end convolutional networks, e.g. Chen et al. [63] who reported the maximum normalised frontal Menpo error to be less than 0.05.

Considering the physical resource usage outlined in Table 4, it is evident that patch-based AAMs can be constructed with significantly less memory requirements. Using the large HELEN dataset as an example, the patch AAM only required 6.3 GB compared to the 26.1 GB of the holistic version. In terms of CPU usage and workload, it is difficult to draw any meaningful comparison between the holistic and patch AAMs. The resources used depended more upon the dataset than the model. If the reader wishes to recreate these models, Table 4 does provide a useful indicator for determining if the model can be created using the reader's own equipment.

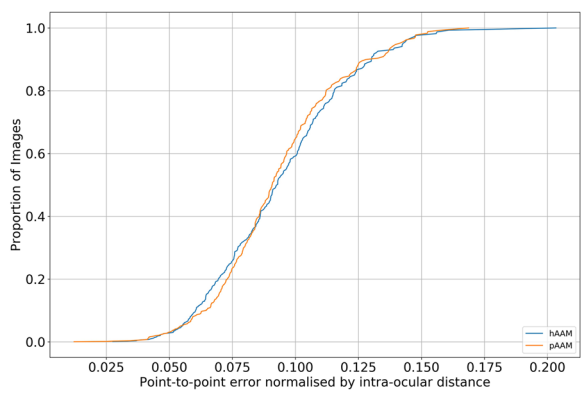

Fig. 15 HELEN cumulative error distributions. Comparing the performance of holistic AAMs and patch AAMs

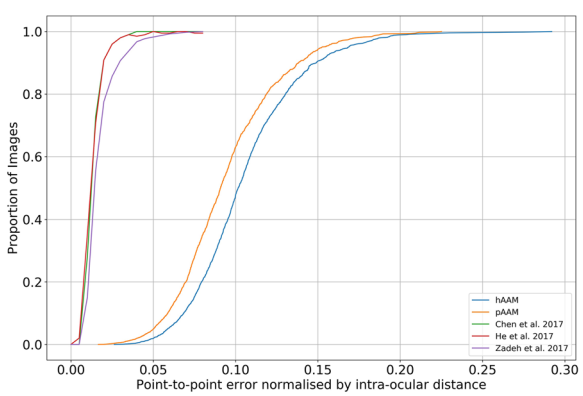

Fig. 17 Menpo (frontal) cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Chen et al. [63], He et al. [67], and Zadeh et al. [61]

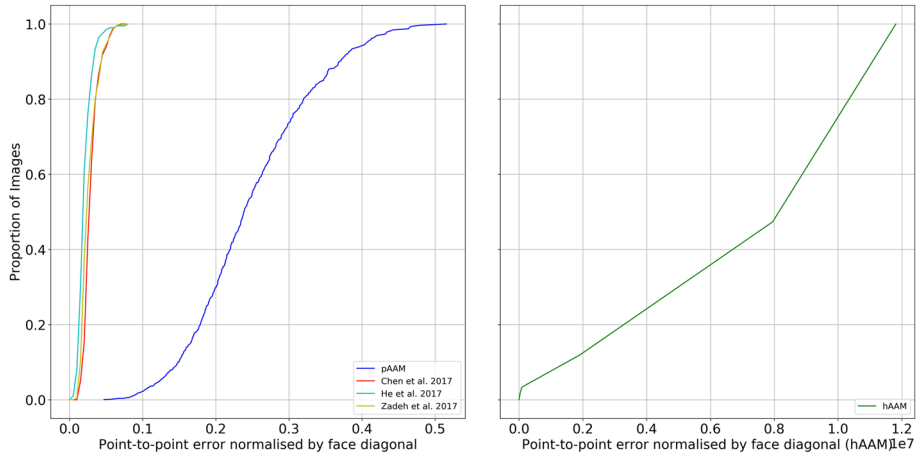

Fig. 18 Menpo (profile) cumulative error distributions. Comparing the performance of holistic AAMs, patch AAMs, and the performance reported by Chen et al. [63], He et al. [67], and Zadeh et al. [61]

2.5.12 Conclusion

In conclusion, this section has provided examples of the 'typical' performance of AAMs using a selection of readily available facial datasets. These performance results reported alongside that of published literature provide the reader with an indication of the improvements made by more recent deep CNN-based methods.

2.6 Stage 5: model training and evaluation

The final stage in the automatic facial landmarking process is training and assessing the defined model, given all of the choices made in stages 1 to 4. Unlike the preceding stages, this training process is not unique for the problem of automated facial landmarking and employs methods well described within existing 'machine learning literature'. As such, a thorough discussion of these methods is considered out of scope for this survey article. If readers are unfamiliar with existing literature, we recommend *Deep Learning* by Goodfellow et al. [56], *Understanding the difficulty of training deep feedforward neural networks* by Glorot and Bengio [73], and the many publications by Yann Le Cun (for practical tips [74] is a useful reference) for an introduction.

Table 4 AAM resources required

Dataset	Model	Time (h:min:s)	Avg # CPUs	Workload (min × CPUs)	Memory (GB)
BioID	hAAM	00:03:06	4.45	14	1
BioID	pAAM	00:06:12	4.70	29	1.1
300W (indoor)	hAAM	00:06:14	5.53	34	2.6
300W (indoor)	pAAM	00:03:09	4.47	15	1.8
300W (outdoor)	hAAM	00:06:14	5.53	34	2.6
300W (outdoor)	pAAM	00:03:09	4.47	14	1.8
300W (combined)	hAAM	00:15:41	5.12	80	4.4
300W (combined)	pAAM	00:09:10	3.79	35	4.3
HELEN	hAAM	03:53:59	6.89	1612	26.1
HELEN	pAAM	00:34:50	6.40	223	6.3
MUCT	hAAM	01:26:56	7.81	679	4.9
MUCT	pAAM	01:36:25	7.77	749	4.6
Menpo (frontal)	hAAM	06:28:44	3048	7.83	16.6
Menpo (frontal)	pAAM	05:44:16	2702	7.85	16.1
Menpo (profile)	hAAM	00:28:56	138	4.76	5.4
Menpo (profile)	pAAM	00:36:44	224	6.09	5.4

3 Conclusions

Throughout this review, we have considered the problem of automatic facial landmarking in its generic form and have investigated the current state of the art within each individual stage of the process. Reflecting on the progress that has been made within this domain, it can certainly be seen that more recent techniques in multi-stage convolutional neural networks have achieved a significant improvement in performance and have contributed to improvements in face detection as well as landmark localisation. With such improvements and given the potential applications for automatic facial landmarking such as that mentioned in the Section 1 one would logically ask 'is the current state of the art good enough?' and 'what is the optimal performance that one could expect from an automatic landmarking system?' In answering the first of these two questions, we must consider the context in which the system will be used. In reasonably controlled environments as those provided within the MULTI-PIE, MUCT, and BioId databases, with minimal variance in expression, lighting and pose, current technology may be more than sufficient. However, for applications where

these conditions would be highly variable such as attempting to classify one's mood in the outdoors using facial expression, current technology may still struggle.

Considering the optimal expected performance, it is our opinion that the current best performing vision systems are still biological. The vision system of humans is capable of accurately identifying facial characteristics given a very high degree of variability in almost all aspects of an image or 'real-life' situations. This differential in performance is demonstrated by Sagonas et al. [10] which compared the landmarking variability of three expert annotators to the automated methods of the 300W competition. The human annotators outperformed the best automatic system by more than a factor of five (Table 3 [10]). In the context of extremely sophisticated applications such as autonomous cars, drones, and robots, a digital vision system with performance comparable to existing biological systems would be of enormous value. If the field is to achieve biological or near-biological performance in automated facial landmarking, we believe that additional progress must be made in all five stages of model construction described within this paper. Redefining the problem objective as *achieving near-biological system performance in facial landmarking*, it is believed that the field will require:

1. A significant increase in the availability of complex 'in-the-wild' data sets with an extremely high sample size, robust ground truth landmarks, and high variability in facial pose, expression, degree of obstruction, and lighting.
2. An improvement in the accuracy of face detection methods as well moving towards methods which describe the outline of the face in more detail, rather than simply specifying the extent to which a face fits within a bounding box.
3. An improvement in modelling methods and learning techniques which are able to capture the characteristics and variance within the datasets in fine detail, without suffering from overfitting.

What is particularly interesting is that in previous years, this list would have included some increase in hardware specifications such as memory or processing power. However, given the major increases in cloud computing and distributed data processing, we do not consider computing resources to be a significant current constraint. Cloud computing providers such as Amazon Web Services and Google Cloud Computing offer configurable instances of CPUs, GPUs with extensible RAM which are readily accessible and at relatively low cost.

Abbreviations

AAM: Active appearance model; AdaGrad: Adaptive gradient algorithm; Adam: Adaptive moment estimation; ASM: Active shape model; CD-Net: Candidate decision network; CE-CLM: Convolutional experts constrained local model; CED: Cumulative error distribution; CEN: Convolutional experts network; CLM: Constrained local model; CNN: Convolutional neural network; CPU: Central processing unit; DAN: Deep alignment network; DA-Net: Dataset across network; FEC-CNN: Fully end to end cascaded convolutional neural network; GPU: Graphical processing unit; hAAM: Holistic active appearance model; HOG: Histogram of oriented gradients; NMRSE: Normalised root mean squared error; OSA: Obstructive sleep apnoea; pAAM: Patch active appearance model; PDM: Point distribution model; RMSE: Root mean squared error; ROI: Region(s) of interest; SVM: Support vector machine

Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their valuable comments.

Funding

No funding has been provided for this work

Availability of data and materials

All data, source code, and other materials used and generated for this article are publicly available under a variety of open source licenses:

- Table 1 contains a list of publicly available facial landmarking datasets, including those used in the experimental components of this review paper.
- The source code for all experimental components of this paper is licensed under GPL 3.0 and can be obtained from GitLab.
- Docker containers with executable forms of the source code are also available via DockerHub
- The image landmarking tool used for the ground truth varaibility study is also available under GPL 3.0 on GitLab.

Authors' contributions

BJ was responsible for the design, research, and writing of the paper as well executing the experimental components. PC provided valuable assistance in planning and preparing the experimental components, reviewing, shaping, and improving the quality of the final manuscript. Both authors read and approved the final manuscript.

Authors' information

Benjamin Johnston is a PhD candidate at the University of Sydney and a Senior Data Scientist at ResMed Ltd. Benjamin's current research interests include computer vision, image processing, pattern recognition, and the application of these technologies in providing improved medical care.
Professor Philip de Chazal is the ResMed Chair in Biomedical Engineering at the University of Sydney. Professor de Chazal leads the University's research and educational activities in biomedical engineering, working across the Charles Perkins Centre and the Faculty of Engineering and Information Technologies.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 27 December 2017 Accepted: 24 August 2018
Published online: 15 September 2018

References

1. T. Wu, P. Turaga, R. Chellappa, Age estimation and face verification across aging using landmarks. IEEE Trans. Inf. Forensic Secur. **7**(6), 1780–1788 (2012). https://doi.org/10.1109/TIFS.2012.2213812
2. T. Devries, K. Biswaranjan, G. W. Taylor, in *2014 Canadian Conference on Computer and Robot Vision*. Multi-task learning of facial landmarks and expression (IEEE, Montreal, 2014), pp. 98–103. https://doi.org/10.1109/CRV.2014.21, http://ieeexplore.ieee.org/document/6816830/
3. G. DL, J. Dusseldorp, H. TA, N. Jowett, A machine learning approach for automated facial measurements in facial palsy. JAMA Facial Plast. Surg. **20**(4), 335 (2018). https://doi.org/10.1001/jamafacial.2018.0030

4. S. Anping, X. Guoliang, D. Xuehai, S. Jiaxin, X. Gang, Z. Wu, Assessment for facial nerve paralysis based on facial asymmetry. Australas. Phys. Eng. Sci. Med. **40**(4), 851–860 (2017). https://doi.org/10.1007/s13246-017-0597-4
5. A. Tabatabaei Balaei, K. Sutherland, P. Cistulli, P. de Chazal, in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. Automatic detection of obstructive sleep apnea using facial images, (Melbourne, 2017), pp. 215–218. https://doi.org/10.1109/ISBI.2017.7950504
6. B. Johnston, A. McEwan, P. de Chazal, in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Semi-automated nasal PAP mask sizing using facial photographs (IEEE, Jeju Island, 2017), pp. 1214–1217. https://doi.org/10.1109/EMBC.2017.8037049, http://ieeexplore.ieee.org/document/8037049/
7. O. Çeliktutan, S. Ulukaya, B. Sankur, A comparative study of face landmarking techniques. EURASIP J. Image Video Process. **13**, 1–27 (2013). https://doi.org/10.1186/1687-5281-2013-13
8. F. Marcolin, E. Vezzetti, Novel descriptors for geometrical 3D face analysis. Multimedia Tools Appl. **76**(12), 13805–13834 (2017). https://doi.org/10.1007/s11042-016-3741-3
9. E. Vezzetti, F. Marcolin, S. Tornincasa, L. Ulrich, N. Dagnes, 3D geometry-based automatic landmark localization in presence of facial occlusions. Multimedia Tools Appl. (2017). https://doi.org/10.1007/s11042-017-5025-y
10. C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces In-The-Wild Challenge: database and results. Image Vis. Comput. **47**, 3–18 (2016). https://doi.org/10.1016/j.imavis.2016.01.002
11. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, in *2013 IEEE International Conference on Computer Vision Workshops*. 300 Faces in-the-wild challenge: The first facial landmark localization challenge (IEEE, Sydney, 2013), pp. 397–403. https://doi.org/10.1109/ICCVW.2013.59, http://ieeexplore.ieee.org/document/6755925/
12. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, in *2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. A semi-automatic methodology for facial landmark annotation (IEEE, Portland, 2013), pp. 896–903. https://doi.org/10.1109/CVPRW.2013.132, http://ieeexplore.ieee.org/document/6595977/
13. Y. Wu, T. Hassner, K. Kim, G. Medioni, P. Natarajan, Facial landmark detection with tweaked convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell (2015). https://doi.org/10.1109/TPAMI.2017.2787130, http://ieeexplore.ieee.org/document/8239860/
14. P. Viola, M. J. Jones, Robust real-time object detection. Int. J. Comput. Vis. **February**, 1–30 (2001). https://doi.org/10.1.1.23.2751
15. N. Dalal, B. Triggs, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol.1. Histograms of oriented gradients for human detection (IEEE, San Diego, 2005), pp. 886–893. https://doi.org/10.1109/CVPR.2005.177, http://ieeexplore.ieee.org/document/1467360/
16. G. Bradski, OpenCV Library (2000). OpenCV.org. Accessed 5 Sept 2018
17. T. Li, W. Hou, F. Lyu, Y. Lei, C. Xiao, in *2016 Sixth International Conference on Instrumentation Measurement, Computer, Communication and Control (IMCCC)*. Face detection based on depth information using HOG-LBP (IEEE, Harbin, 2016), pp. 779–784. https://doi.org/10.1109/IMCCC.2016.92, http://ieeexplore.ieee.org/document/7774889/
18. F. Song, X. Tan, X. Liu, S. Chen, Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. Pattern Recognit. **47**(9), 2825–2838 (2014). https://doi.org/10.1016/j.patcog.2014.03.024
19. A. Suleiman, V. Sze, An energy-efficient hardware implementation of HOG-based object detection at 1080HD 60 fps with multi-scale support. J. Signal Process. Syst. **84**(3), 325–337 (2016). https://doi.org/10.1007/s11265-015-1080-7
20. R. Vaillant, C. Monrocq, Y. L. Cun, Original approach for the localisation of objects in images. IEE Proc. Vis. Image Signal Process. **141**(4), 245–250 (1994). https://doi.org/10.1049/ip-vis:19941301
21. M. Osadchy, Y. Le Cun, M. L. Miller, Synergistic face detection and pose estimation with energy-based models. J. Mach. Learn. Res. **8**, 1197–1215 (2007). https://doi.org/10.1007/11957959
22. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE. **86**(11), 2278–2323 (1998). https://doi.org/10.1109/5.726791. http://arxiv.org/abs/1102.0183
23. L. Haoxiang, L. Zhe, S. Xiaohui, J. Brandt, H. Gang, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. A convolutional neural network cascade for face detection (IEEE, Boston, 2015), pp. 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170
24. S. Yang, P. Luo, C. C. Loy, X. Tang, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. WIDER FACE: A face detection benchmark (IEEE, Las Vegas, 2016). https://doi.org/10.1109/CVPR.2016.596, http://ieeexplore.ieee.org/document/7780965/
25. X. Sun, P. Wu, S. C. H. Hoi, Face detection using deep learning: an improved faster RCNN approach. Neurocomputing. **299**, 42–50 (2017)
26. R. Girshick, J. Donahue, T. Darrell, J. Malik, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Rich feature hierarchies for accurate object detection and semantic segmentation (IEEE, Columbus, 2014), pp. 580–587. https://doi.org/10.1109/CVPR.2014.81, http://ieeexplore.ieee.org/document/6909475/
27. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
28. R. Uijlings, A. van de Sande, T. Gevers, M. Smeulders, et al., selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154 (2013)
29. S. Bell, C. L. Zitnick, K. Bala, R. Girshick, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks (IEEE, Las Vegas, 2016), pp. 2874–2883. http://doi.org/10.1109/CVPR.2016.314, http://ieeexplore.ieee.org/document/7780683/
30. S. Zhan, Q. Q. Tao, X. H. Li, Face detection using representation learning. Neurocomputing. **187**, 1–8 (2015). https://doi.org/10.1016/j.neucom.2015.07.130
31. Q. Q. Tao, S. Zhan, X. H. Li, T. Kurihara, Robust face detection using local CNN and SVM based on kernel combination. Neurocomputing. **211**, 98–105 (2016). https://doi.org/10.1016/j.neucom.2015.10.139
32. D. Wang, J. Yang, J. Deng, Q. Liu, FaceHunter: A multi-task convolutional neural network based face detector. Signal Process. Image Commun. **47**, 476–481 (2016). https://doi.org/10.1016/j.image.2016.04.004
33. S. Zafeiriou, C. Zhang, Z. Zhang, A survey on face detection in the wild: Past, present and future. Comp. Vision Image Underst. **138**, 1–24 (2015). https://doi.org/10.1016/j.cviu.2015.03.015
34. M. Kawulok, M. E. Celebi, B. Smolka, *Advances in Face Detection and Facial Image Analysis*. (Springer, Cham, 2016). https://doi.org/10.1007/978-3-319-25958-1. http://link.springer.com/10.1007/978-3-319-25958-1
35. D. E. King, Dlib-ml: A Machine Learning Toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009). https://doi.org/10.1145/1577069.1755843
36. A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Incremental face alignment in the wild (IEEE, Columbus, 2014), pp. 1859–1866. https://doi.org/10.1109/CVPR.2014.240, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909636
37. J. Deng, Q. Liu, J. Yang, D. Tao, M3 CSR: Multi-view, multi-scale and multi-component cascade shape regression. Image Vis. Comput. **47**, 19–26 (2015). https://doi.org/10.1016/j.imavis.2015.11.005
38. G. J. Edwards, C. J. Taylor, T. F. Cootes, in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. Interpreting face images using active appearance models (IEEE Comput. Soc, Nara, 1998), pp. 300–305. https://doi.org/10.1109/AFGR.1998.670965, http://ieeexplore.ieee.org/document/670965/
39. T. F. Cootes, G. J. Edwards, C. J. Taylor, active appearance models. IEEE Trans. Pattern. Anal. Mach. Intell. **23**(6), 681–685 (2001)
40. T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active shape models-Their training and application. Comp. Vision Image Underst. **61**(1), 38–59 (1995). https://doi.org/10.1006/cviu.1995.1004
41. J. C. Gower, Generalized procrustes analysis. Psychometrika. **40**(1), 33–51 (1975). https://doi.org/10.1007/BF02291478
42. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. Philos. Mag. Ser. 6. **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720
43. G. J. Edwards, C. J. Taylor, T. F. Cootes, in *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition*. Learning to identify and track faces in image sequences (IEEE, Comput. Soc, Nara, 1998), pp. 260–265. https://doi.org/10.1109/AFGR.1998.670958, http://ieeexplore.ieee.org/document/670958/
44. T. F. Cootes, C. J. Taylor, *Statistical models of appearance for computer vision. Technical report*. (University of Manchester, Manchester, 2008). http://www.face-rec.org/algorithms/#AAMmodels.pdf

45. T. Baltrusaitis, P. Robinson, L. P. P. Morency, in *2013 IEEE International Conference on Computer Vision Workshops*. Constrained local neural fields for robust facial landmark detection in the wild, (2013), pp. 354–361. https://doi.org/10.1109/ICCVW.2013.54. http://ieeexplore.ieee.org/document/6755919/
46. D. Cristinacce, T. F. Cootes, in *Procedings of the British Machine Vision Conference*. Feature detection and tracking with constrained local models (British Machine Vision Association, Edinburgh, 2006), pp. 95–19510. https://doi.org/10.5244/C.20.95
47. J. Peng, L. Bo, J. Xu, Conditional neural fields. Adv. Neural Inf. Process. Syst. **9**, 1–9 (2009)
48. J. Yan, Z. Lei, D. Yi, S. Z. Li, in *Proceedings of the IEEE International Conference on Computer Vision*. Learn to combine multiple hypotheses for accurate face alignment, (2013), pp. 392–396. https://doi.org/10.1109/ICCVW.2013.126
49. X. Xiong, F. De La Torre, Supervised descent method and its applications to face alignment. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 532–539 (2013). https://doi.org/10.1109/CVPR.2013.75
50. D. G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94. http://arxiv.org/abs/0112017
51. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst., 1–9 (2012). https://doi.org/10.1016/j.protcy.2014.09.007. http://arxiv.org/abs/1102.0183
52. L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, R. Fergus, in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, vol. 28, ed. by S. Dasgupta, D. Mcallester. Regularization of neural networks using DropConnect, (2013), pp. 1058–1066. http://jmlr.org/proceedings/papers/v28/wan13.pdf
53. A. Krizhevsky, *Learning multiple layers of features from tiny images*. (University of Toronto, 2009), pp. 1–60. https://doi.org/10.1.1.222.9220. http://arxiv.org/abs/arXiv:1011.1669v3
54. B. Graham, Fractional Max-Pooling. International conference on learning representations, 1–10 (2014). http://arxiv.org/abs/1412.6071
55. D.-A. Clevert, T. Unterthiner, S. Hochreiter, in *International Conference on Learning Representations 2016 (ICLR 2016)*. Fast and accurate deep network learning by exponential linear units (elus) (ICLR, San Juan, 2015), pp. 1–14. https://arxiv.org/abs/1511.07289
56. I. Goodfellow, Y. Bengio, A. Courville, in *Deep Learning*. Convolutional Networks (MIT Press, Cambridge, 2016), pp. 330–372. http://www.deeplearningbook.org
57. S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. The Menpo Facial Landmark Localisation Challenge: A Step Towards the Solution (IEEE, Honolulu, 2017). https://doi.org/10.1109/CVPRW.2017.263. http://ieeexplore.ieee.org/document/8014997/
58. E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, in *2013 IEEE International Conference on Computer Vision Workshops*. Extensive facial landmark localization with coarse-to-fine convolutional network cascade (Institute of Electrical and Electronics Engineers Inc., Sydney, 2013), pp. 386–391
59. H. Fan, E. Zhou, Approaching human level facial landmark localization by deep learning. Image Vision Comput. Online J. **47**, 27–35 (2016). https://doi.org/10.1016/j.imavis.2015.11.004
60. J. Yang, Q. Liu, K. Zhang, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Stacked hourglass network for robust facial landmark localisation, (2017)
61. A. Zadeh, T. Baltrusaitis, L. P. Morency, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Convolutional experts constrained local model for facial landmark detection, (2017)
62. Y. Sun, X. Wang, X. Tang, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Deep convolutional network cascade for facial point detection, (2013), pp. 3476–3483. https://doi.org/10.1109/CVPR.2013.446
63. X. Chen, E. Zhou, Y. Mo, J. Liu, Z. Cao, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Delving deep into coarse-to-fine framework for facial landmark localization, (2017)
64. C. Zhang, Z. Zhang, in *2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014*. Improving multiview face detection with multi-task deep convolutional neural networks, (2014), pp. 1036–1041. https://doi.org/10.1109/WACV.2014.6835990
65. A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation. Eur. Conf. Comput. Vis., 483–499 (2016). https://doi.org/10.1007/978-3-319-46484-8. http://arxiv.org/abs/1603.06937
66. W. Wu, S. Yang, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Leveraging intra and inter-dataset variations for robust face alignment, (2017)
67. Z. He, M. Kan, J. Zhang, X. Chen, S. Shan, in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. A fully end-to-end cascaded CNN for facial landmark detection (IEEE, Washington, D.C, 2017), pp. 200–207. https://doi.org/10.1109/FG.2017.33
68. Z. He, J. Zhang, M. Kan, S. Shan, X. Chen, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Robust FEC-CNN: A high accuracy facial landmark detection system (IEEE, Honolulu, 2017). http://doi.org/10.1109/CVPRW.2017.255. http://ieeexplore.ieee.org/document/8014989/
69. J. M. Saragih, S. Lucey, J. F. Cohn, Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. **91**(2), 200–215 (2011). https://doi.org/10.1007/s11263-010-0380-4
70. M. Kowalski, J. Naruniec, T. Trzcinski, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Deep alignment network: A convolutional neural network for robust face alignment, (2017)
71. C. Sagonas, S. Zafeiriou, Facial point annotations (2017). https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/ Accessed 18 May 2017
72. S. Milborrow, T. E. Bishop, F. Nicolls, in *Proceedings of the IEEE International Conference on Computer Vision*. Multiview active shape models with SIFT descriptors for the 300-W face landmark challenge, (2013), pp. 378–385. https://doi.org/10.1109/ICCVW.2013.57
73. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. Proc. 13th Int. Conf. Artif. Intell. Stat. (AISTATS). **9**, 249–256 (2010). https://doi.org/10.1.1.207.2059
74. Y. LeCun, G. Montavon, G. B. Orr, K. R. Mülle, Y. LeCun, in *Neural Networks: Tricks of the Trade*. Lecture Notes in Computer Science, vol. 7700, ed. by .G. Montavon, G. B. Orr, and K. R. Müller. Efficient BackProp (Springer, Berlin, Heidelberg, 2012), pp. 9–48. https://doi.org/10.1007/978-3-642-35289-8
75. R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*. Multi-PIE, (2008), pp. 1–8. https://doi.org/10.1109/AFGR.2008.4813399
76. S. Milborrow, J. Morkel, F. Nicolls, The MUCT landmarked face database. Pattern Recog. Assoc. S. Afr. (2010)
77. K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, in *Second International Conference on Audio and Video-based Biometric Person Authentication*, vol. 964. XM2VTSDB: The extended M2VTS database (AVBPR, Washington, D.C, 1999), pp. 965–966
78. S. Zafeiriou, M. Pantic, G. Chrysos, G. Trigeorgis, J. Deng, J. Shen, 2nd facial landmark localisation competition - The Menpo benchMark (2017)
79. M. Koestinger, P. Wohlhart, P. M. Roth, H. Bischof, in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, (2011). https://doi.org/10.1109/ICCVW.2011.6130513. http://ieeexplore.ieee.org/document/6130513/
80. A. Kasinski, A. Florek, A. Schmidt, The PUT face database. Image Process. Commun. **13**(3-4), 59–64 (2008)
81. M. Fink, R. Fergus, A. Angelova, Caltech 10,000 web faces (2007). http://www.vision.caltech.edu/Image_Datasets/Caltech_10K_WebFaces/
82. O. Jesorsky, K. J. Kirchberg, R. W. Frischholz, in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Robust face detection using the hausdorff distance (Springer, 2001), pp. 90–95
83. V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, in *European Conference on Computer Vision - ECCV 2012, Lecture Notes in Computer Science), vol. 7574 LNCS*. Interactive facial feature localization (Springer, Berlin, 2012), pp. 679–692. https://doi.org/10.1007/978-3-642-33712-349. http://link.springer.com/10.1007/978-3-642-33712-349
84. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, (2008)